SISK, CECILIA, Ph.D. Effects of Ignored Subpopulations' Growth Trajectories on Estimates of School Value-Added Scores. (2018).
Directed by Dr. Robert Henson. 192 pp.

School value-added scores classify schools into performance categories that are linked to rewards and sanctions. Because value-added scores claim to measure the schools' effectiveness on student growth, inferences of the quality of services provided are made. However, the widespread use of these scores has not yet been sufficiently supported by research as a sound accountability index, particularly when it pertains to its accurate interpretation and its ensuing appropriate use for high stakes decisions. Research shows that several factors can change the classification of schools such as methodology, constructs used, variables used and others. In order to add to the body of evidence of whether the inferences derived from value-added scores can be supported, this research will investigate the effects of un-accounted for latent subpopulations, LEP and student SES at level-1 and school SES at level-2 on the classification of schools' value-added scores and its precision estimates in multilevel data utilizing the multilevel growth mixture model and multilevel linear growth model. This research found that the number of schools identified for special treatment were cut in half when value-added scores were extracted from a multilevel growth mixture model in conjunction with the specification of school SES at level-2, in comparison to a multilevel linear growth model without school SES at level-2. Particularly, the value-added scores' magnitude were less extreme for the more homogeneous schools, the very high SES and the very low SES schools. In addition, precision estimates were improved as well. This suggests that using the methodology that sanctions the larger number of schools would be premature because there are other factors that can affect the value-added scores estimates.

EFFECTS OF IGNORED SUBPOPULATIONS' GROWTH TRAJECTORIES ON ESTIMATES OF SCHOOL VALUE-ADDED SCORES

by

Cecilia Sisk

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2018

Approved by

_____
Committee Chair

APPROVAL PAGE

This dissertation written by Cecilia Sisk has been approved by the following committee of the Faculty of The Graduate School at The University of North Carolina at Greensboro.

Committee Chair_____

Committee Members_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

LIST OF TABLES.....vi

LIST OF FIGURES.....viii

LIST OF ABBREVIATIONS.....ix

CHAPTER

I. INTRODUCTION.....1

1.1 School Accreditation.....1
1.2 Use of Standardized Scores.....3
1.3 Value-Added Model as a Potential Solution.....5
1.4 VAM Limitations.....5
1.4.1 Heterogeneity in Teacher/School Effect Estimates.....7
1.5 Study Proposal.....8

II. LITERATURE REVIEW.....10

2.1 Value-Added Models.....11
2.1.1 Value-Added Models Overview.....12
2.1.2 VAMs Strengths and Limitations.....15
2.2 More Advances in Multilevel Modeling.....20
2.3 Educational Research with Advanced Multilevel Modeling.....24
2.3.1 Latent Covariates and the General Mixture Model.....25
2.3.2 The Latent Growth Curves and Growth Mixture Models.....28
2.3.3 Example of the Use of MLGMM in an Educational Setting.....29
2.3.4 MLGMM in VAM.....30
2.4 Algebraic Descriptions of MLM.....33
2.4.1 Algebraic Description of Longitudinal Multilevel Model.....33
2.4.2 Algebraic Description of Latent Growth Model.....34
2.4.3 Algebraic Description of Multilevel Latent Growth Model.....35
2.4.4 Algebraic Description of Growth Mixture Model.....38
2.4.5 Algebraic Description of Multilevel Growth Mixture Model (MLGMM).....40
2.5 Purpose of this Study.....46

III. METHODOLOGY 49

3.1 Cohort Description 49
3.2 Characteristics of my Empirical Study 54
3.3 Characteristics of Individual (Level-1) Data 55
3.3.1 Latent Classes Identification 55
3.3.2 Level-1 Estimation 57
3.4 Data Study Framework 58
3.5 Student SES as Level-1 Covariate 65
3.6 Adding LEP as Level-1 Covariate 67
3.7 Characteristics of Cluster (Level-2) Data 70
3.7.1 School Variability 71
3.7.2 Cluster Effects 72
3.8 Model Fitting 74
3.8.1 Analysis of Results of Model Fitting 75
3.9 School Effects Analysis 77
3.9.1 Evaluation of Value-Added Scores Classification: Disagreement Rates 79
3.10 Summary of Methods 81

IV. MAIN STUDY RESULTS 82

4.1 Model Identification 1: Longitudinal Multilevel Model Level-1 82
4.1.1 Basic Analysis 84
4.1.2 Unconditional Model (No Covariates Model) 85
4.1.3 Add Covariates - FRL 88
4.1.4 Add Covariates FRL and LEP 89
4.2 Model Identification 1: Two-Level Conventional MLM 91
4.2.1 Unconditional Two-Level MLM 92
4.2.2 Conditional Two-Level MLM (School SES) 98
4.3 Model Specification 2: Growth Mixture Model Analysis Results Level-1 101
4.3.1 Growth Mixture Modeling with Latent Classes: Models A and B 102
4.3.2 Multinomial Logistic Regression Parameterization 103
4.3.3 Reading Scores Unconditional Model Comparisons 103
4.3.4 Conditional Models Comparisons 104
4.3.5 Growth Mixture Full (FRL and LEP) Model Results 107
4.4 Model Specification 2: MLGMM Analysis Results Level-2 111
4.4.1 Unconditional Two-Level MLGMM 113
4.4.2 Conditional Two-Level MLGMM 118

4.5 School Value-Added Estimates 125
4.5.1 School Value-Added Formulation Specifications 125
4.5.2 Standardized Value-Added Scores and Thresholds Specifications 127
4.5.3 School Value-Added Estimates ANCOVA Results 130
4.5.4 School Value-Added Precision Estimates 132
4.5.5 School Value-Added Classification Disagreement Rates 137

V. DISCUSSION 140

5.1 Model Identification Process with MLM 144
5.2 Model Identification Process with MLGMM 145
5.2.1 Information Criteria Performance 145
5.2.2 MLGMM Convergence and Interpretation of Effects 146
5.3 School Value-Added Effects 147
5.3.1 Results on School Value-Added Classification 147
5.3.2 Results on School Value-Added Magnitude Effects 148
5.3.3 Precision of Estimates 149
5.4 Limitations of the Research 149
5.5 Future Directions 151

REFERENCES 165

APPENDIX A. SCHOOL VALUE-ADDED SCORES 173

LIST OF TABLES

Page

Table 1. Grades 3 through 6 Cohort Sample Description 50

Table 2. Percent of Students Not-Proficient from 2011-2014 53

Table 3. Summary Statistics Standardized Reading Scores 53

Table 4. Growth Parameter Settings for the Four Latent Classes 57

Table 5. Correlation Matrix for Reading Scores for Times 1, 2, 3 and 4 84

Table 6. MLM One-Level Within Level Estimates 87

Table 7. MLM Fit Indices for One-Level 91

Table 8. MLM Two-Level Within Level Estimates 94

Table 9. MLM Two-Level Between Level Estimates 97

Table 10. MLM Fit Indices for Two-Level 98

Table 11. Restrictions per Model Specification 102

Table 12. GMM Model Comparisons Class Fit 104

Table 13. MLGMM Level-1 Model Comparisons Fit 105

Table 14. MLGMM Class Membership Unconditional 106

Table 15. MLGMM Class Membership FRL only 106

Table 16. MLGMM Class Membership FRL and LEP 107

Table 17. GMM Within Estimates FRL and LEP 108

Table 18. GMM Multinomial Estimates FRL and LEP 111

Table 19. MLGMM Multinomial Estimates without School SES 114

Table 20. MLGMM Within Estimates without School SES 116

Table 21. MLGMM Between Estimates without School SES 117

Table 22. MLGMM Class Membership without School SES 118

Table 23. MLGMM Multinomial Estimates with School SES 121

Table 24. MLGMM Within Estimates with School SES 122

Table 25. MLGMM Between Estimates with School SES 123

Table 26. MLGMM Class Membership with School SES 124

Table 27. MLGMM Fit with and without School SES 125

Table A1a. Unstandardized Value-Added Scores without School SES 173

Table A1b. Unstandardized Value-Added Scores with School SES 174

Table A2a. Standardized Value-Added Scores without School SES 175

Table A2b. Standardized Value-Added Scores with School SES 176

Table A3. Estimates of Methodology Difference by School SES (high to low) 177

Table A4. Classes Mixture Composition by School SES 178

Table A5a. School-Level Analysis: System Error Rates without School SES 179

Table A5b. School-Level Analysis: System Error Rates with School SES 180

LIST OF FIGURES

Page

Figure 1. Graphical Description of GMM 40

Figure 2. Graphical Description of MLGMM 45

Figure 3. Advanced MLMs 46

Figure 4. Graphical Description of Growth Profiles 48

Figure 5. Four-Class MLGMM with LEP, FRL and School SES 113

Figure 6. Reading Grades 3-6 HP, LP, S and PLP 119

Figure 7. Reading Grades 3-6 Latent Classes Initial Status (C_I) and Growth Rate (C_S) 120

Figure A1a. Unstandardized Value-Added Scores without School SES 181

Figure A1b. Unstandardized Value-Added Scores with School SES 182

Figure A2a. Standardized Value-Added Scores without School SES 183

Figure A2b. Standardized Value-Added Scores with School SES 184

Figure A3a. Ranked Standardized Value-Added Scores without School SES 185

Figure A3b. Ranked Standardized Value-Added Scores with School SES 185

Figure A4a. Value-Added Scores SD Estimates without School SES 186

Figure A4b. Value-Added Scores SD Estimates with School SES 187

Figure A5a. Value-Added Scores SE Estimates without School SES 188

Figure A5b. Value-Added Scores SE Estimates with School SES 189

Figure A6a. Ranked Value-Added SE without School SES 190

Figure A6b. Ranked Value-Added SE with School SES 190

Figure A7. Standardized Value-Added Scores for all Models 191

Figure A8. Standardized Value-Added Scores SE for all Models 192

LIST OF ABBREVIATIONS

AIC	Akaike Information criteria
AIC3	Modified Akaike information criteria
AICc	Second order bias corrected AIC
AYP	Adequate yearly progress
BIC	Bayesian information criteria
BICB	Bayesian information criteria with a cluster number for sample size adjustment factor
CFI	Comparative fit index
CS	Cluster size
DPI	Department of Public Instruction
EVAAS	Education value-added assessment system
FRL	Free and reduced lunch
GMM	Growth mixture model
HP	High performing
LC	Latent class
LCV	Latent class variable
LEP	Limited English proficient
LGCM	Latent growth curve model
LGM	Latent growth model
LN	Natural logarithm
LP	Low-performing
M	Mean

MAR	Missing at random
ML	Maximum likelihood
MLGMM	Multilevel Growth Mixture Model
MLLGM	Multilevel latent growth model
MLM	Multilevel longitudinal model
MLM	Multilevel model
NC	North Carolina
NCLB	No Child Left Behind
PeLP	Chronically low performing
PLP	Permanently low performing
RMSEA	Root mean square error of approximation
S	Striving
SD	Standard deviation
SABIC	Sample size adjusted BIC
SEM	Structural equation modeling
SES	Socio-economic status
SRMR	Standardized root mean square residual
TAP	Teacher Advancement Program
TLI	Tucker Lewis index
TVAAS	Tennessee value-added assessment system
U.S.	United States
VAM	Value-added model
VM	Value-added modeling

CHAPTER I

INTRODUCTION

I present an alternative methodological approach to isolate the contribution of schools to students' learning gains as captured by standardized Reading tests. The literature on value-added models has been mainly focused on teachers' effect on students. However, the same models can be utilized to estimate schools' contribution to students' achievement. For this reason, the literature mentioned can also support my work. In those cases for which the literature is pertinent for teachers and schools, I will refer to it as "teacher/school". The remainder of the paper is as follows: in Chapter 1, I present a brief introduction. In Chapter 2, I present my literature review on VAMs and the current methodological approaches. In Chapter 3, I describe the data and details of the methodological framework. In Chapter 4, I present my results. In Chapter 5, I present my discussion.

1.1 School Accreditation

Annual standardized student testing is a pervasive piece of the U.S. K–12 educational system, but this has not resolved the age-old issue about whether and how policymakers should use the test results in accountability systems. During the 1970s, many states expanded student testing and adopted minimum competency exams that students had to pass to graduate from high school. In the 1980s, states added school report cards which were reporting point-in-time snapshots of average school achievement (Harris, 2008). This trend toward test-based school level accountability accelerated in the 1990s with state policies such as school grades, reconstitution, takeovers, and other incentives (Harris & Herrington, 2007). NCLB appears to have cemented school-level test-based accountability as a key element in the national education

strategy, but this is a broad strategy that allows for a wide variety of policies regarding the use of standardized test scores (Harris, 2008). The current educational accountability policy holds the system, including schools, principals and teachers, responsible for the academic advancements of the students. However, the policy's assessment tools include only some crude indicators (e.g. percent of students who are proficient or school means on End-of Grade tests) to measure teacher/school impact on students. The current goal of schools and teachers is to increase learning for all students, reduce achievement gaps, and improve system efficiency. However, there is an emphasis on only assessing test-based outputs with minimal local context (e.g. school funding). The standard reference indicators are simple, incomplete, and based on "theory-of-action", thus do not necessarily help to achieve meaningful consequences (Braun, 2005).

The Theory-of-Action calls for justification for imposing a particular accountability system with the promise that it will accomplish the desired goals. Too often, the Theory-of-Action is stated in simplistic terms and ignores other (less desirable) behavioral responses that the accountability system may elicit such as identifying students who exhibit very low performance (Siegel & Filardo, 2011). The "theory-of- action" behind NCLB is that test-based accountability will improve the productivity of our nation's public school system by using indicators based on summaries of individual test scores, setting stretch goals, focusing on each sub-population, providing supplementary educational services and threatening sanctions (e.g. school choice, school restructuring) (Braun, 2005; Harris, 2011). NCLB contains an implicit assumption that any technical flaws in the indicators are of secondary concern and that incentives (and consequences) will result in greater attention to appropriate goals and more effective administrative and pedagogical strategies.

1.2 Use of Standardized Scores

Overall student achievement scores have two main purposes: 1) creating summative assessments of effectiveness that determine who is performing well; and, 2) creating (ideally) a basis for incentives that produce student growth (Gordon, Kane & Staiger, 2006; Hanushek & Rivkin, 2006; Kane, Rockoff & Staiger, 2008, Goldhaber 2008; Harris, 2015). Currently, school accountability has a strong empirical component based primarily on a test score-based criterion of continuous improvement (AYP). To evaluate AYP, a school's score must be computed for all students in a grade, as well as for various subgroups, the proportions meeting a fixed standard (e.g. proportion of students who are proficient in state standardized assessments, also known as "Percent Proficient") and then compare these proportions with those obtained in the previous year. Several observers have pointed out the problems arising from making AYP judgments about schools or teachers based on the concept of an absolute standard (Linn, 2004; Linn, 2007; Neal, 2010; Glazerman, 2011). Specifically, students entering with a higher level of achievement will have less difficulty meeting the proficiency standard than those who enter with a lower level of achievement, since the former may have already met the standard or may be very close to it. As a result, some students must make little or no progress to contribute to the school's target. Indicators based on "Percent Proficient" are unsound, especially when used to track changes and gaps.

There is evidence in the literature of other external factors outside of schools or teacher control that affect student academic achievement. The fundamental problem in holding schools/teachers accountable for student achievement is that education are jointly produced by schools, families, and communities (Hanushek, 1979; Rothstein, 2004; Ermisch & Pronzato, 2012; Ladd, 2012; Coley & Baker, 2013; Garcia, 2015, Morsy & Rothstein, 2015). Students' SES is arguably the strongest predictor of their educational outcomes, an observation dating at

least as far back as Coleman (1966). More current research shows that low SES children are 1.3 SDs lower than high SES children in their Kindergarten entry math skills (Duncan & Magnuson, 2011), 0.8 SDs below in reading, 0.4 SDs lower in persistence in completing tasks (Garcia, 2015), 0.70 SDs below in teacher ratings of attention skills and 0.25 worse in terms of teacher-reported antisocial behavior (Deming 2009). Garcia and Weiss (2015) demonstrated that Black and Hispanic ELL children start kindergarten with the greatest disadvantages in math and reading, due largely to links between minority status and social class. Since such differences occur before students enter school, their source must be family, community and other factors beyond school control. It is therefore no surprise that schools serving White students from middle and high income families are far more likely than a high-minority, high-poverty school to be among a state's top-third on achievement tests (Harris, 2007; Palardy,2013).

These facts pose difficulties for school/teacher accountability systems, whose expressed goals are to measure and reward school/teacher performance. If school/teacher performance measures substantially reflect non-school/non-teacher contributors to student success, as is the case with NCLB and typical state school report cards, then genuine improvements in school/teacher performance are not appropriately assessed using higher performance measures, leaving schools with only a weak incentive to improve, a perverse incentive to prefer the most socio-economically advantaged students in the classrooms (Harris, 2007; Harris, 2011) and a need to minimize the presence low of performers (Figlio,2012; de la Torre & Gwynne, 2009; Glazerman, 2011; Ballou & Springer, 2015). In this sense, school accountability based on such misleading performance measures is not only unfair to the schools but may also be unfair to the students.

1.3 Value-Added Model as a Potential Solution

VM has drawn significant attention as a way to solve some of the problems with standardized scores discussed above by trying to isolate the contribution of schools/teachers. VM has interested individuals at different levels (school superintendents, policy makers, NC DPI, government officials and researchers) of the education debate is accountability based on how much "value-added" teachers and schools contribute to student achievement. One attraction of VAMs (Goldhaber, Harris, Loeb, McCaffrey, & Raudenbush, 2015; Harris, 2011; McCaffrey, Lockwood, Koretz, & Hamilton, 2003) is that this approach to accountability diff ers in a critical way from the AYP provisions of the NCLB Act. This accountability indicator is based on student gains and it addresses some of the problems with the current NCLB regulations: Gain scores broaden the empirical basis for evaluation, gains are more weakly correlated with student characteristics when compared to student status, this indicator reduces the impact of differences between cohorts and its use enhances the perception of fairness. The basic logic is: if each student's achievement is measured every year, then, in trying to determine each school's performance, we can take into account where students started at the beginning of each year and therefore indirectly account for the family and community factors that contribute to achievement. This approach differs from typical school report cards and from NCLB, neither of which account for where students start. Another advantage that the VM approach has over simpler indicators is that it permits the inclusion of multiple sources of bias that could affect VAM-derived estimates such as student level characteristics, or school/teacher characteristics if they are known to be relevant and/or able to be obtained.

1.4 VAM Limitations

VAMs have their own problems (e.g., unaccounted variability) and the traditional VAMs poorly address the issues of bias due to non-random effects (Kane, McCaffrey, Miller, & Staiger,

2013; Rothstein, 2010; Raudenbush, 2013). Even advocates of test-based accountability acknowledge that measuring teacher/school contributions to student test scores is difficult. VM still cannot isolate the teacher/school effect on student growth because many assumptions must hold in order to interpret value-added measures as truly causal (i.e., treating them as statistically unbiased estimates of teacher/school eff ectiveness) teacher/school contributions to student learning gains (Braun, 2005; Harris, 2009; Rothstein, 2010; Raudenbush; 2013). One important assumption is that after factors such as students' past performance and some student characteristics (e.g., ethnicity or student SES) are included in the model any extra sources of variability in the data are random with one normal distribution. In other words, the student growth patterns within a school or a teacher vary randomly and the school or teacher estimates are the unbiased representation of school or teacher contribution to student learning gains.

One problem with this argument is that many of the most common student level variables used to account for the lack of random assignment of students to teachers or schools are distal predictors of student's achievement (e.g., bias at level-1). In addition, the bias of student scores (e.g., bias at level-2) due to sorting or students' selection to schools (e.g., low SES students tend to attend low performing schools) cannot be completely removed by traditional student level characteristics because key variables that account for student learning gains are never obtained (Ladd 2012; Rothstein 2004; Coley and Baker 2013). Some evidence shows that there is an indirect link between SES and outcomes through the statistically significant associations between economic (dis)advantage and multiple factors also related to educational results. These factors include the environment in which a child grows up (neighborhood factors and family characteristics), a child's participation in early childhood programs, the quality of those programs, and even the type and quantity of instructional and motivational activities that parents engage in with their children that affect child development. All of these associations are

significant, and all help better explain educational growth. However, none of these variables, which represent the real opportunities that children have been given—and the needs that they have, are assessed or included in VAMs (Ladd 2012; Rothstein 2004; Coley and Baker 2013).

1.4.1 Heterogeneity in Teacher/School Effect Estimates

The fact that school systems cannot or do not record relevant variables which influence student learning gains suggests that many other external factors may produce potential systematic variability (as opposed to random variability) in a VAM. This systematic variability in school estimates created by the inability to explicitly control for relevant variables may result in biased school/teacher estimates of contribution to student growth. This bias is because results from the fact that the usual variables (e.g. ethnicity or student SES) used to describes students' characteristics in VAM do not capture the variability caused by the presence of multiple subpopulations in the data; the traditional VAM cannot accommodate the presence of this variability in producing estimators across schools/teachers. If the school's effect on a group of students is assumed to be from multiple distributions, the group is called heterogeneous. I assume that any actual heterogeneity represents unknown or uncontrolled sources of level-1 (also known as student observations) variability in the VAM. I argue here that a potentially large source of heterogeneity resides in the variation of groups of students sharing similar but not explicitly observed or obtained background characteristics, thus these groups contain a mix of subpopulations. Each subpopulation has its own distribution. It seems possible that this is quite common to heterogeneous subpopulations, particularly among low SES students (Garcia, 2015). Some scholars have suggested that while value-added measures may be valid in some overall sense, they may not be valid estimators for teachers with certain types of students (Harris & Anderson, 2013; Jackson, 2012) since the presence student growth patterns within a school/teacher resulting from heterogeneous distributions violates the VAM assumption of level-

1 random variation with one normal distribution. Researchers conducting simulation studies (Yumoto, 2011; Asparouhov,2009) using VAMs have been able to account for heterogeneity of distributions of individuals' learning gains through the use of latent classes (i.e. two latent classes), each of which its own growth profile, (i.e. starting point and growth rate).

1.5 Study Proposal

Greater discernment is needed when looking at children's growth by subgroups. To effectively identify the performance and needs of highly diverse groups of children, analysts must first categorize types of students using common underlying characteristics and identify more homogenous subgroups. One possible approach involves identifying and grouping students of similar performance growth profiles to overcome the inability to explicitly control for relevant variables on learning gains (Asparouhov,2009). More advanced multivariate methods have been developed to model heterogeneity in the data if unspecified effects differ systematically across schools. To address the omitted variables problem, I propose to estimate a standard VAM with the addition of a latent variable indicating the growth performance profile of the student with a MLGMM (Asparouhov, 2009), in addition to controlling for school SES at level-2 to address the student selection problem.

In this paper, I make two contributions to the broader teacher/school quality literature. First, I advance Sanders' method by presenting a new approach to estimating value-added effects with a MLGMM using empirical data. MLGMM has the capability to capture types of students based on their growth profiles in the form of latent classes and at the same time produce model-based VAM estimates. As mentioned before, schools are unable or unwilling to obtain information on relevant predictors of student learning gains; for this reason, traditional VAMs must be based on distal predictors of student performance potentially leaving uncontrolled sources of variability. I will try to capture that systematic variability through latent variables.

Motivated by Yumoto (2011), who suggested two types of students (High and Low Performers), I propose the existence of four different types of student growth performing types (which will be discussed in more detail later in this document): High Performers, Strivers, Low Performers and Persistently Low Performers. My second contribution is to advance the VAM research and its usefulness in policy to improve teaching and learning. I posit that traditional VAMs yield more extreme teacher/ school estimates, particularly for schools with a large number of either high SES students or low SES students. More specifically, school estimates with a large number of high SES students tend to be positive and very high in magnitude and schools estimates with a large number of low SES students tend to be negative and very high in magnitude. However, in simulation studies comparing schools/teachers controlling for student growth profile, there are less extreme teacher/school estimates (Yumoto, 2011) thus I expect that after accounting for student growth profile effects at level-1 and school SES at level-2 the estimates of school/teacher effects will be less extreme than estimates from the traditional VAM.

My results could potentially influence policy. School effects on student achievement may be considerably less pessimistic for schools with large numbers of low SES students when the analytical method is able to account for key subpopulations in the data unlike the traditional VAM. These results would lead decision makers away from closing some schools deemed "low performing schools" based on more pessimistic traditional school value-added estimates. Note that I focus on policy-relevant persistent school effects that can be reasonably attributed to schools (rather than variability that may reflect systematic variability due to family or community resources not specified in the model). Glazerman (2011) has suggested that when policymakers rely on flawed measures of school performance, they risk closing schools that are better prepared to work with challenging populations.

CHAPTER II

LITERATURE REVIEW

Currently, school/teacher accountability includes a strong empirical component that is based primarily on a test score-based criterion of continuous improvement (AYP). Inferences about schools/teachers contributions to student learning gains are still based only on a small set of externally mandated criteria represented as simple means or the percent of students who are proficient on standardized tests. However, indicators based on "Percent Proficient" are flawed, particularly when used to track changes and gaps. Since random assignment of students to teachers/schools is usually not practiced, simple means (or "Percent Proficient") of class/school achievement test scores are biased by many factors (e.g. parental influences) other than teacher influences that affect student learning. Any system that will fairly and reliably assess the influence of teachers/schools on student learning must partition teacher effects from other factors. "Percent Proficient" measures of teacher/school effects are confounded with differences among schools, family and community contexts. In addition, education is a cumulative process. The educational resources students receive early in life affect their academic success later in life, but it is impossible to explicitly measure the complete range of resources that students receive at any given time, let alone in past years (Harris, 2015).

The impact of employing test based indicators for educator evaluation depends on the operating characteristics of the accountability system (consequential validity) as a whole. Accountability practices are systematically valid if they contribute to the improvement of one or more of the goals of access, quality, equity and efficiency for an education system (Braun & Kanjee, 2006; Ballou, 2015). Judging systematic validity requires a comprehensive effort in data

collection and data analysis. Current efforts (e.g. "Percent Proficient"), if they are even attempted, suffer from evidential asymmetry, because they mainly focus on test scores and top-down regulation as a crude tool for organizational improvement (http://www.dpi.state.nc.us/accountability; Corcoran, 2012).

2.1 Value-Added Models

The VM (Sanders & Rivers, 1996) represents attempts to circumvent many of the problems associated with the use of student achievement data in assessment of school systems, schools and teachers through reliance on the scale scores that indicate gains students make from year to year, regardless of the points at which the students enter the classroom. By focusing on measures of academic gain, each student serves as his or her own control. In other words, each child can be thought of as a "blocking factor" that enables the estimation of school/teacher effects on the academic gain with the need for few, if any, variables. Student growth provides complementary descriptions of what is happening with respect to one aspect of student learning, for instance, a student's starting achievement level (i.e. intercept) and amount of growth (i.e. slope) in a given number of years.

Also, if deemed necessary, student characteristics can be added to the VAMs because student characteristics provide additional descriptions of what is happening with respect to other characteristics of student learning. Student characteristics are included to adjust for confounding from external factors not under the control of teachers/schools by adding concomitant co-variables (e.g. student SES, class size, funding, curricular innovations and others) as needed. Overall, the potential attraction of VM methods is that these approaches may indirectly account for unknown (or unobtainable) external factors even when non-random assignment of students to schools/teachers occurs (Sanders, 1994). Consequently, VM may produce a somewhat "level playing field" that produces fair (unbiased) comparisons across teachers/schools.

2.1.1 Value-Added Models Overview

The term VAMs and their application in educational settings date at least as far back to Hanushek (1979) and Boardman and Murnane (1979). In an attempt to partition the teacher and school effects from the partial confounding with initial status's students ability level, Millman (1981) suggested the use of linear model techniques of analysis of covariance and ordinary multiple regression with the intent to adjust for differences that exist among students to enable a fairer evaluation of teachers. In 1984, Dr. William Sanders and Dr. Robert McLean, statisticians from the University of Tennessee, published a working paper on the use of student achievement data as a basis for teacher assessment. They utilized three years of gain scores from Knox County students' performance on the California Achievement Test in grades 3 through 5 and developed a statistical system of analysis. Later in 1991, when the Education Improvement Act was adopted, Sanders' model was incorporated as part of the Tennessee educational accountability system (TVAAS). When the NCLB legislation was established, Sander's model was integrated in several states in the U.S. including NC (where it is known by the acronym EVAAS (SAS Institute Inc, 2000-present). The VAM is commonly implemented in longitudinal multi-level modeling frameworks to capture the contribution of higher-level effects such as schools/teachers (level-2 or cluster/group effect) on the student's achievement and/or improvement (level-1 or individual level) over time (Sanders & Rivers, 1996). The multilevel model is used to account for the interdependence of individuals within the same group/cluster and model the effects of both individual-level (i.e. student growth rate or slope and starting point or intercept) and group-level variation on an outcome simultaneously (Pollack, 1998).

In general, VM refers to quantitative approaches that estimate the contributions of teachers, schools or programs to students' achievements, taking into account the differences in prior achievement and (perhaps) other student characteristics. The predicted score is a

counter factual (i.e. an estimate of the outcome after exposure to the average unit) which is usually compared to a relative criterion or an absolute criterion, so different VAM models yield different rankings of units (Braun, 2005; Harris, 2011; Harris; 2015). In the case where the comparison is a relative criterion, VAMs yield estimated teacher effects defined as deviations from the average teacher/school in the district and teacher/school whose contributions are determined to be significantly different from the average can be identified for special treatment. In the case where the comparison is an absolute criterion, teacher/school are compared to each other by some criteria such as teacher/school SES (often defined as the percentage of students who receive free and/or reduced lunch). In order to capture a student's gain in different subject matters (e.g., Math or Reading), the basic VAM defines the score of a student at the end of the school year as the sum of three components: the district average for that grade and year, the class (teacher) effect, and other systematic and non-systematic variations. Thus, the essential difference between the student's score and the average score in the district is attributable to a cluster effect (e.g., teacher or school) plus the combined contributions of unspecified variations, including measurement error. It is assumed that the cluster effect is the same for all the students in the classroom/school and attributable to the teacher of the classroom/school, therefore, it is referred to it as classroom/school effect. The identification of the cluster effect with teacher effectiveness conflates two separate ideas: 1) endowing a statistical quantity (cluster effect) with a causal interpretation; and, 2) attributing the causal contribution of the cluster entirely to the teacher/school (Braun, 2005; Harris, 2008; Harris and Sass, 2007b; Jackson and Bruegmann, 2009; Clark, Martorell & Rockoff, 2009). When the student moves to the next year and the next grade, the model then has four components: district average for that grade and year, teacher effect for that year, teacher effect from the previous year, and other systematic and random variations (Braun, 2005).

The assumption here is that the teacher/school effect for the previous year persists undiminished into the current year and that the components of the unspecified variations in the two years are unrelated to each other. Finally, if we subtract the first year score from the second year score, we obtain the gain made by the student. According to the model, this must be the sum of the average gain for that grade in the district, the teacher effect of the second year teacher, and the two error terms; that is, ignoring the error terms, the teacher effect in the second year is the difference between the gain experienced by the student in that year and the average gain in the district for that same year. It is possible to add equations for the data from subsequent years. Sanders (1996) uses the term "layered model" to capture the notion that the data from each succeeding year are added to those from the previous years. In a typical application, students may contribute as many as five years of data. Moreover, student gains in different subjects are included in the EVAAS model (SAS Institute Inc., 2000), with each subject and year assigned its own equation. With the initial version of the VAM, Sanders argued that there is no need to include student characteristics in the model. His rationale was that, while there are substantial correlations between these characteristics and the current level of achievement, the correlations of these characteristics with gains are essentially zero. Also, he assumes that controlling for a student's previous achievement accounts for the impact of all of past student resources that may contribute to learning gains such that, when comparing student progress from year to year, many of the student characteristics considered to influence (for example) a student's fourth-grade achievement are the same as those influencing her third-grade achievement. The change in a student's score will cancel out these external factors and reveal only the impact of changes since the third-grade test, with the year of fourth-grade instruction being the most obvious (Corcoran, 2010).

Later, the model was revised to include student level fixed effect variables because economists found that including student fixed effects in their VAMs enabled them to address the lack of random assignment of students to teacher/school (Rosthein, 2012). Fixed effect variables are variables that do not change over time and may include some student characteristics such as gender, ethnicity or SES. They are also called time invariant manifest variables and are the most common type of variables specified in the VAMs. This version of the model is potentially more realistic than the previous version which only utilized past achievement.

The current model version represents the conditional average rate of achievement growth over all the years for which there are student scores in the database and they are conditional in the sense that fixed effects control for external factors that might influence student learning in any given year that are largely outside school/teachers' control. However, only specifying the model with student level fixed effects does not rule out the possibility that other relevant variables, not specified in the model (therefore part of level-1 variation), may also account for learning gains. However, VAMs do contain an assumption that these omitted variables (or unobtainable variables) are randomly distributed among teachers/schools. In other words, all unaccounted variability is randomly distributed with one distribution within teachers/schools. If the level-1 random variability assumption is violated, bias may be introduced into the teacher/school value-added measures even when student fixed effects are included in the model (Harris and Sass, 2007a; Harris, 2008; Kane and Staiger, 2008).

2.1.2 VAMs Strengths and Limitations

One of the strengths of VAMs is the focus on gains while accounting for the fact that students began the year at very different levels (Corcoran, 2010). Controlling for student's previous achievement accounts to some extent for the impact of past student resources which may have been employed to achieve that gain. However, controlling for past achievement does not

account for the fact that students are assigned to schools/teachers based on other observed student characteristics that may also be related to students' subsequent achievement. There is evidence that assignment of students to teacher/schools is correlated with factors (e.g. teacher experience, etc.) that are sometimes related to teacher/school value-added score (Clotfelter et al., 2005; Harris, 2008; Monk, 1987; Gamoran, 2010; Feldman, 2009). Other scholars (Feng, 2005; Chudowsky, 2007; Hattie, 2012) have found that students are assigned to teachers partly based on students' discipline problems. Harris and Sass (2005), McCaffrey, Sass, and Lockwood (2010), and Aaron (2012) showed that findings regarding teacher value-added scores are quite different when relying solely on controlling for past achievement as opposed to when some fixed student characteristics are added in the model.

From a statistical standpoint, isolating the impact of schools/teachers from other external factors is a problem of non-random assignment (otherwise is impossible to assure the random variability assumption) of students to teachers and teachers to schools. According to statistical theory, the ideal setting for obtaining proper estimates of teacher/school effectiveness is a school system in which, for each grade, students are randomly grouped into classes/schools, and teachers in that grade are randomly allocated to those classes/schools. Roughly speaking, randomization levels the playing field for all teachers/schools in that each teacher/school has an equal chance of being assigned to any class/school; then, determining that the average student growth associated with a particular teacher/school is significantly greater than the district average would be credible evidenc for that teacher's/school's relative effectiveness (Clotfelter et al., 2005, Gamoran, 1986; Oakes, 1985; Ogbu, 2003, Harris, 2010).

Adding fixed effects to the model ameliorates to some extent the lack of random assignment of students to schools/teachers but time invariant manifest variables in the model are still deficient proxies used to statistically remove some of the systematic error from the

achievement gain measure. There remain potential numerous other external factors that have not been accounted for and that could be more informative in predicting student achievement gain (Garcia, 2015). Some of these variables may be time variant variables. Time variant variables are variables that change over time such as the English proficiency of a child who is learning English as her second language. Many of the factors that matter most, such as family resources, parental involvement, greater out of school support and student ability, are difficult, if not impossible, to quantify (Corcoran, 2010). Many of these student-level factors are random events, while others systematically affect teachers/schools from year to year (Corcoran, 2010). Comparing units on the basis of current scores or gain scores ignores the confounding of unit effects by other relevant factors that are differentially distributed across units (students).

The claims of proponents of the VAM that teachers/schools are the main source of variation in student gains are questionable given that fixed effect variables are imperfect indicators of learning gains. Estimates based on VAMs are also descriptions and using such descriptions for the purpose of accountability implicitly assumes that they are accurate indicators of school (or teacher) effectiveness. Assuming that statistically estimated effects are the true and only representation of teacher/school effectiveness on student achievement can be dangerous when the data for the analysis come from an observational study and not a randomized experiment. Assuming that a statistical effect is the same as teacher /school effectiveness is a fundamental problem of causal inference (Holland, 1986). Despite its theoretical appeal, isolating a school/teacher's unique contribution is very difficult. A host of factors that have nothing to do with teacher effectiveness and are not randomly distributed among the students in the classroom can affect student academic growth. Models using only manifest time invariant variables cannot fully eliminate the selection bias caused by non-random pairing of students and schools (or teachers) nor can they replace the key variables that explain student gain. Simply said, VAM

proponents assume any statistical bias is too small to worry about. Unfortunately, most of the assumptions made are not directly testable. Thus, the credibility of the causal interpretations, as well as the inferences and actions based on them, must depend on the plausibility of such assumptions (Harris, 2008; Harris & Sass, 2005 & Todd & Wolpin, 2003). In order to understand the pitfalls in interpreting VAM results as indicators of school or teacher effectiveness, we must examine the inference chain more carefully. For instance, some evidence indicates the presence of mediating mechanisms between SES and outcomes on student achievement gains through multiple factors (e.g. parental engagement and early childhood programs) also related to education results (Coley and Baker, 2013; Ladd, 2012; Rothstein, 2004). Identifying these factors could result in the detection of subgroups of students of differing array of performance profiles that cannot be captured simply by students' SES.

In addition, in many contexts, attempts to attribute achievement gains to individual teachers/schools may not be practical, particularly in middle and high school, when students receive instruction from multiple teachers. To assume that none of these teachers' effects "spill over" into other coursework may be a strong and unrealistic assumption. Indeed, Koedel (2009) found that reading achievement in high school is influenced by both English and Math teachers. Learning may not simply occur in the rigid process assumed by current VAMs. Consequently, teachers rewarded or punished for their value-added-assessed effect on a student's gains (or lack thereof) may, in part, be rewarded or punished based on the teachers with whom they work. This possibility certainly runs counter to the intended goal of value-added assessment.

Statistical models irrespective of their complexity are always simplifications of the data they represent: when summarizing the growth of students in classrooms/schools, the mean value is a (very simple) model that collapses over the distribution (Miles & Shevlin, 2000), thereby masking features such as whether the distribution is multimodal, whether there are outliers, and

so forth. For example, if one classroom has more high SES students, and this fact was not incorporated into the estimate of the mean (naturally resulting in a more complex summary of the classroom's weight), then (assuming an uniform average effect of SES) one classroom could appear to have a higher average growth. A model which does not take the SES of students into account produces bias. Similarly, when applying the VAM approach to estimating teacher/school effects on student performance, assumptions are made that might affect the estimates or their interpretation.

Scholars describe the result of not specifying relevant variables in the model (including time manifest variables and latent classes) as *hidden heterogeneity*. An example of this occurs when the data contains subpopulations with their own distributions distinct from the overall population and these subpopulations are not specified in the model. The potential consequences of hidden heterogeneity include biased and inconsistent cluster level estimates (Yumoto, 2011). The direction of the bias depends on the sign of estimates in the model, the nature of the omitted variables and the degree of correlation between the omitted variables and the remaining variables. There is no good diagnostic test for hidden heterogeneity. Most diagnostics tests assess for correlation between independent variables and the error term, but such correlation may result from many problems including measurement error. The existence of omitted variables cannot here fore be determined conclusively (Yumoto, 2011).

As such, hidden heterogeneity due to the lack of identification of student subpopulations creates hurdles for the evaluation design as well as the analysis and interpretation of the model results. In the context of VAM-derived school effects, systematic variability, or heterogeneity among the students (due to the lack of specification of subpopulations in the model) represents a clear violation of the assumption that student growth patterns are from a single homogeneous distribution and thus exist as a normal distribution with fixed mean and variance within a cluster

(school or teacher). Thus, if student growth patterns are from a mixture of distributions (i.e., are heterogeneous) within clusters, the variation attributable to the un-modeled distribution inflates the variation at a higher level in the model (e.g., at level-2, the level of the classroom/teacher/school), causing mis-estimation and even misinterpretation of the school's effect. When it is assumed that students are from a single homogeneous distribution, or that the cluster's effect on the students varies only randomly, any actual heterogeneity represents unknown or uncontrolled sources of variability in the system – violating the VAM assumption.

I argue that traditional manifest variables (e.g., student SES and ethnicity) are not sufficient to capture the relevant variability of students' learning gains and that there is systematic variation in the level-1 of VAMs resulting in biased cluster level (teacher or school) estimates. These typically-included variables do not completely capture the diversity of the performance profiles among students (which may result from past student resources). As a result, heterogeneous subpopulations are not explicitly identified in the model. Heterogeneity in this instance means the systematic variability that remains in level-1 VAMs due to the inability of traditional models to capture these different subpopulations of performance profiles with only manifest student variables (e.g. ethnicity), since each subpopulation contains its own distribution which may not correspond to one overall random distribution (as assumed by VAMs).

2.2 More Advances in Multilevel Modeling

Here I detail the results of my literature review for the methodology used by traditional VAMs and other new advances in methodology that could potentially be useful in VAMs. The traditional VAM uses a basic MLM in which repeated observations over time (test scores) are nested within students, and students are nested within teachers or schools (also known as the cluster, between or group level) (Verbeke and Molenberghs, 2000). This model accounts for variability within students within teachers/schools and it also accommodates for any sources of

bias at the student or cluster level (by adding covariates at each level). However, this model is still deficient in successfully isolating cluster level effects because the model assumes that relevant sources that explain students' learning gains have been explicitly identified. In this sense, the model is restrictive because data collected must be available to account for this potential systematic variability. In many cases, school districts do not have the expertise to even know what variables should be collected, let alone have the manpower to collect the data. As a result, school databases only contain variables that are very easy to collect such as gender or ethnicity.

The multilevel model is generally used to account for the interdependence of individuals within the same group and model the effects of both individual-level and group-level variation (i.e., heterogeneity) on an outcome simultaneously (Pollack, 1998). In multilevel models, the time variable is considered level-1, student characteristics are considered level-2 and the teacher/school characteristics (which I refer to as cluster level) are considered level-3. Consistent with that, I refer to each nested structure in these terms as I further discuss these models (e.g. time nested within students nested within schools). However, as I review the more complicated models this terminology will change because the more complex models utilize a SEM framework, wherein the student and the repeated observations are all treated as observations for the student. For that reason, student and time levels are considered to be at the same level, which I will refer to as level-1 (also known as the individual, student or within-subject), and the school/teacher level is considered level-2 (also known as the group or between-subject), also referred to as the cluster level.

Researchers have developed and refined methods for multivariate models, adding gradually more complex mechanisms for modeling variability in the data under study. One of these developments is multilevel modeling, which has become increasingly widespread in

educational research. In 1972, Lindley and Smith presented the first multilevel model – in the Journal of the Royal Statistical Society (Lindley & Smith, 1972) - which they developed to accommodate variability across individuals that confounded precise estimation of the level-2 (cluster) parameters of interest; random effects can include variation in group-level (level-2) parameters (e.g., group level mean/intercept), or degree of level-2 mean deviation from the overall group-level (level-2) mean (Nezlek & Zyzniewski, 1998). The random effect represents an additional level of analysis, so that regression coefficients become random variables; with observations nested within, e.g., individuals (for whom a single constant regression coefficient would be estimated). Verbeke and Molenberghs (2000) describe the specification of random effects as the second of a two stage modeling method ("general linear mixed modeling"); considering their "stages" as levels corresponds to a multilevel model. Burstein, Linn, and Capell (1978) utilized multi-level data analysis to accommodate the presence of heterogeneity in regression estimators across classrooms within a single study population. The treatment of data as explicitly hierarchical, with observations at one level (e.g., at the individual level) nested within other levels (e.g., the cluster level) depends critically on how the levels and hierarchy are described and defined (see Kreft et al., 1995). To maintain generality, we refer to this type of model as an MLM.

Considering three levels where scores are nested within students that are nested within schools enables more precise (less biased) estimates of student level (level-1) effects of post-test scores (Raudenbush & Bryk, 2002) under some circumstances; that is, by planning for and accommodating the heterogeneity arising from specific features in the data, the effect of variability on the estimates can be minimized. For example, if students within a school are more homogeneous (random variation is lower) than the overall student population, accounting for the clustering within the data (e.g., modeling students as if they are nested within a school) aids in the

allocation of variation in an outcome measure across different levels (student, school). In this example, accommodating the lower level variability within this school improves precision and may reduce bias in growth parameter (level-1) estimates across individuals and/or across schools. Accounting for level-1 variability may also reduce the bias in fixed effects estimates at level-2 and level-1 because these parameters might be affected by some students' characteristics.

Before performing an analysis the analyst must identify the relationship(s) in the data and what it is in the data that needs to be modeled. A specification that exists in different subpopulation of students, and a further specification that they are not randomly distributed at the cluster level, will inform both study design and data collection, shaping the research design or hypothesis. By specifying in the model the nature of the model structure (key predictors at each level of analysis, including identification of subpopulations), the analysis will yield better estimates (random and fixed) at each level which in turn will provide better information in how to make effective decisions on policy development and resource allocation necessary to support each subgroup of students' learning gains. However, planning for heterogeneity in the data subpopulations is also critical to the research process in order to support a valid interpretation of results from any statistical analysis. The importance of directly modeling this variability is reflected in both empirical studies (e.g., Lazarsfeld & Henry, 1968; Goodman, 1974; Clogg & Goodman, 1985; Muthén & Shedden, 1999; Muthen & Muthen & Nagin, 1999; Jo, 2002; Kreuter & Muthén, 2008; Henry & Muthén, 2010) and methodological development work (e.g., Lazarsfeld, 1950; Quandt, 1958; Quandt & Ramsey, 1972; Goodman, 1974; Titterington, Smith & Makov, 1985; Verbeke & Lesaffre, 1996; Bartholomew & Knott, 1999; McLachlan & Peel, 2000; Muthén, 2001; Vermunt & Magidson, 2002, Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004; Bollen & Curran, 2006; Asparouhov & Muthén, 2008;). Methodological work has benefitted from and expanded to accommodate and model the influence of both observed

(manifest) and unobserved (latent) variables in the estimation and interpretation of multivariate statistical analysis (Loehlin, 1998). Software applications such as *MPlus* (Muthén & Muthén, Ver 7.4, 2016), Latent Gold (Vermunt & Magidson, Ver 5.1, 2016), Lisrel (Joreskog & Sorbom, Ver 9.1, 2014) and EQS (Bentler, Ver 6.3, 2016) have both increased and supported the capacity of investigators to consider and analyze manifest and unobserved contributors to the variability in their data (Feldman, Masyn & Conger, 2009; Henry & Muthén, 2010; Jo, 2002; Kreuter, Yan, & Tourangeau, 2008; Marsh et al., 2009; Preacher, Zyphur, & Zhang, 2010; Schaeffer et al., 2006). Several estimates from statistical models can vary depending on whether manifest and/or latent variables are modeled (Hancock & Lawrence, 2006; Muthén & Asparouhov, 2009) – and particularly whether these are modeled appropriately or not (Chen et al., 2010; Palardy & Vermunt, 2010). Since the estimates can vary in relation to these features, so, too can the inferences based on those estimates.

These advances in modeling latent variables can be useful in VAMs because latent classes may be able to account for subpopulations in the data that are normally ignored owing to the fact that the typical manifest variables have no satisfactory explanatory power on learning gains. Some of these advances have been incorporated in MLMs by adding latent variables (inferred from the data) at the student level (level-1) to identify subgroups of students with distinct performance profiles (intercept and slope). Some of these models extend the category known as latent growth models, which is basically a longitudinal MLM. These extensions of latent growth models with added latent class are known as the MLLGM, the GMM and the MLGMM.

2.3 Educational Research with Advanced Multilevel Modeling

In this section I describe the practical utility of using more complex models in different research settings such as medicine, business and particularly in education and how these

advanced models (GMM, MLLGM and MLGMM) can also be applicable to VAMs. Researchers have employed growth curve, mixture, and multi-level models and their recent amalgamations in educational research within the last decade. A common thread of this type of research is to recognize latent classes from growth trajectories that are both qualitative and statistically distinct. Accordingly, the results become more informative in the sense that specific strategies supporting effective instruction (e.g., interventions) can be formulated for each group of students, whether academic (e.g., alternative instruction), behavioral/psychological (e.g., behavioral intervention), or social (e.g., individual counseling). I begin my discussion by describing how researchers identifying latent covariates and general mixture models were effective in finding distinct subpopulations in the model not identified by the manifest variables; next I discuss how later latent covariates were incorporated in latent growth curves to model change over time. Finally, I describe how incorporating latent class variables in the MLM (latent variable identification, growth over time and multi-level modeling) framework can be a logical expansion of VAMs.

2.3.1 Latent Covariates and the General Mixture Model

Muthén and Asparouhov (2009) found that the conventional multilevel model was insufficient to yield precise estimates of level-2 effects; their solution was to utilize latent covariates, inferred from the data, because none of the manifest covariates had any explanatory power. For my work, the latent covariate used to account for the variability at level-1 in Muthén and Asparouhov's (2009) analysis represents a latent class. Vermunt and Magidson (2002) describe a latent class as some factor causing "…some of the parameters of a postulated statistical model <to> differ across unobserved subgroups," (p. 175) where categories of subgroups of this unobserved or latent categorical variable make up the levels of the LC. An LC is therefore a subgroup indicator, similar to a covariate, but it is latent and must be inferred from data. An example of a mixture model, first noted by Lazarsfeld (1950), includes the classification of

applicants into subgroups (e.g., acceptance and rejection groups for uniformed services recruits) built from a set of dichotomous responses on a questionnaire (see also Lazarsfeld & Henry, 1968); that is, the classification of applicants was not based on any observed data, rather the latent (unobserved) classes into which the applicants were sorted were inferred based on their dichotomous responses.

An example of the development and growing support of the capacity of investigators to consider and analyze both manifest and unobserved contributors to heterogeneity is a family of methods called "mixture models" (Muthén, 2002). Verbeke and Molenberghs (2000) refer to a "mixed" model as a regression that includes both random and fixed effects. However, in the more general context (as described in Muthén, 2002), mixture models are a type of statistical method used to conduct an analysis while simultaneously examining if there is more than one sub-population (e.g., at least two subgroups with different distributions) in the data (Muthén et al, 2002; Vermunt & Magidson, 2002).

Mixture models (in this more general sense) have been applied in research domains as diverse as organization (Lazarsfeld, 1950; Kreuter & Muthen, 2008; Shaeffer et al., 2008), education (Dayton, 1991; Muthen et al., 2003; Muthen & Asparouhov, 2009; Palardy & Vermunt, 2010; Muthen, 2004; Asparouhov & Muthen, 2008), and medicine and epidemiology (Croudace et al., 2003; Boscardin et al; 2008). In each case, some analytic method (e.g., linear regression) is the objective, but subpopulations in the data may warrant the use of different regression features. One of the most general mixture models can be defined as an analysis that includes the search for latent subpopulations while simultaneously estimating statistical models including several causal effects, a process beyond straightforward multiple regression. For example, multilevel, structural equation, growth, and the combination of these types of modeling approaches fall under "general mixture models" (see Bartholomew, 1987; Muthén, 1989; Muthén, 2001; Skrondal & Rabe-

Hesketh, 2004; Vermunt & Magidson, 2002). Latent class analysis and finite mixture modeling (McLachlan & Peel, 2000) are technically subsumed within mixture modeling, as they are very specific types of mixture models. This most general formulation of mixture models, which we refer to as "the general mixture model approach" comprises models ranging from simple estimation of a latent class, through less complex models with simultaneous latent class or finite mixture evaluation, to more complex modeling such as latent growth plus latent class/finite mixture combinations.

General mixture mixed models can be used both to identify differential patterns of growth in a group of students and simultaneously to detect the subgroups within the study population for which targeted interventions (e.g., different types of instruction) can be tailored. In educational research, manifest variables such as SES (low/high) are often important covariates, but these should not be confused with LC variables. Less general models (e.g., latent class or finite mixture models) cannot serve these purposes because the primary focus of the less general models is to identify the latent class from the set of observed categorical or continuous variables, instead of permitting the identification of such classes from estimates derived from other simultaneous analyses (Goodman, 1974; Muthén, 2000; Muthen& Muthen & Nagin, 1999). The LC analysis is a valuable analytic method in research where the identification of latent classes is the primary focus. For my work, however, the latent classes represent a complicating feature of the estimation (of school effect), introduced with the intention of reducing bias, and are not an end in themselves.

Exemplifying this potential, Muthén and Asparouhov (2009) used a multilevel mixture model, instead of the conventional multilevel model, where subgroups of students were identified within the latent variable "student type" with levels "fast learner" or "slow learner". This student level (level-1) LCV accounted for the heterogeneity of subpopulations in level-1 residual variance

for which observed covariates or the conventional multilevel model did not account; the mixture model that included this LCV also identified effects which were estimated at the school level (level-2), ultimately changing the estimated effects of covariates at both student and school levels, and leading to different interpretations of parameter estimates than were supported by the conventional two-level model. They also tested for the presence of an LC at the school level and found that, although such a level-2 LC could be identified, it had a very limited impact upon the estimation or interpretation of other parameters. Muthén and Asparouhov's (2009) example showed the importance of thorough investigation of heterogeneity in variance at each level and in particular, that the conventional multi-level model will not always suffice to limit bias and optimize precision of estimates.

2.3.2 The Latent Growth Curves and Growth Mixture Models

Just as hierarchies in data led to multivariate methodological developments such as the multi-level model, individual effects in intercepts and slopes of repeated measures datasets led to the development of the latent growth curve model or growth/growth curve model(Preacher, Wichman, MacCallum, & Briggs, 2008). The purpose of growth models is to model change over time with particular emphasis on the variability in starting points (i.e., intercepts) and change over time (i.e., growth/slope). A latent growth curve mixture model or GMM is an extension of the LGCM. The idea behind the GMM is to allow further examination – and estimation – of the heterogeneity of growth trajectories explainable by latent classes. For example, there may be groups of students with distinctive growth trajectories that cannot be explained well by one set of slopes, intercepts, and their correlations. As noted earlier, accounting for heterogeneities in data is critical to support valid interpretation of results from statistical analysis. The inclusion of slopes and intercepts (growth curve modeling) as multiple levels (multi-level modeling), plus

identification of important covariates such as student type (mixture modeling) are united in the estimation underlying the GMM.

The multilevel extension of GMM was introduced in and has been applied to education (Muthén & Asparouhov, 2009; Palardy & Vermunt, 2010). The Muthén and Asparouhov (2009) example outlined above can be generalized to other educational outcomes like the evaluation of teacher/school effectiveness – which would typically be estimated using a VAM (Sanders & Rivers, 1996). VAM is actually a special case of the GMM; suggesting that growth/growth mixture modeling is a natural tool for estimating the development of student capabilities over time – as well as other effects (e.g., teacher and school) that could be – and may need to be shown to be – contributing to students' growth.

2.3.3 Example of the Use of MLGMM in an Educational Setting

Muthén and Asparouhov (2009) identified the importance of accounting for heterogeneity attributable to an LCV in the context of the MLM framework. Muthén and Asparouhov (2009) applied the multilevel mixture model to simulated data and real data to demonstrate the use and utility of mixture modeling in educational contexts. First, they used a conventional MLM with a single (manifest) level-1 predictor (student-level SES) then added a level-1 latent class covariate (i.e., low and high achievers). Muthén and Asparouhov (2009) showed the different conclusions derived from regression with and without consideration of the latent class when comparing results from a conventional multilevel model against those of a multilevel mixture model.

Muthén and Asparouhov (2009) showed that the effect of student-level covariates can affect the interpretation of the results from a conventional multilevel regression, since the student level latent class variable interacted with the school-level predictor. They also showed that, in the presence of a substantive LCV at the student level especially when the class levels interact with

the covariate, the interpretation of the results will depend on the latent class membership at level-1 and the value of the school-level covariate (i.e., at level-2).

Muthén and Asparouhov (2009) found two specific impacts on estimates and inferences, as compared to the conventional MLM could be derived from the mixture model: 1) Estimates of level-2 effects were inflated in the conventional MLM compared to the mixture model; and, 2) The effects of predictors were significantly different between the conventional and mixture model (these effects were attenuated in the mixture model as compared to the conventional MLM estimates). This supports the importance of modeling the level-1 heterogeneity with latent classes in order to avoid reaching the wrong conclusion by inflating the effect of covariates.

Based on their exploration of the simple regression, conventional MLM, and MLGMM models and their respective fits to the data, in addition to the differing results and inferences supported under each analysis, Muthén and Asparouhov (2009) stated that level-1 heterogeneity in the form of latent classes is mistaken for level 2 heterogeneity in the form of the random effects that are used in conventional two-level regression analysis.

2.3.4. MLGMM in VAM

Two of the papers described earlier, Muthén and Asparouhov (2009) and Palardy and Vermunt (2010), have several important implications for VAM in terms of the correct – MLGMM – analytic approach. An LCV, representing student performance and development, has been identified by two independent groups of researchers (Chudowsky et al., 2007; Lazarus et al., 2010). Both groups identified a subgroup of students that persistently performs at the lowest level. Students are known to be heterogeneous in their performance and their development (Chudowsky, Chudowsky, & Kober, 2007; Lockwood & McCaffrey, 2007; Lazarus et al., 2010), but they may also fall into more predictable (latent) classes that can complicate estimation with growth curve modeling – particularly if this predictable source of variability is ignored (as

reported by Muthén and Asparouhov, 2009). Students who chronically perform at a low level over time have been characterized as PLP students (Chudowsky et al., 2007; Lazarus et al., 2010). These students start off, and remain, at a low performance level over time, and are often distinct from students who start off at a higher level and remain at that level over time as well as from those who start higher or lower and exhibit change over time.

In their study of student types, Lazarus et al. (2010) identified two groups of low performing students, LP and PLP (see also Chudowsky et al., 2007). LP students were defined to be those who scored at the 10th percentile or lower on the state wide standardized test in one of the previous three years. PeLP students were those who scored at the 10th percentile or below on the statewide standardized test for all three years. Those students identified as PeLP were not performing so badly overall that they were eligible to take the alternate form of assessment (i.e., a test for students who are in the special education program), but their performance suggested that the regular achievement tests were simply too difficult for them. Lazarus et al. discovered two demographic (manifest) variables that tended to characterize the PeLP student type: they were more likely to be minorities, and more likely to be receiving free or reduced lunch (a proxy variable for low SES). Although these trends were observed for the manifest demographic variables, neither was statistically significantly predictive of belonging to the PeLP student type. As Palardy and Vermunt (2010) suggested, predictor variables or covariates should not be included for exploration of latent class variables in MLGMM due to the potential interaction between them which may obscure the identification of latent classes. Together with the PeLP results of Lazarus et al. (2010), indicating that manifest covariates are not sufficient, or sufficiently explanatory, the results and recommendations by Palardy and Vermunt (2010) suggest that an LCV – based on slopes and intercepts – may be a more efficient and effective method of identifying students in this class. Palardy and Vermunt (2010) and Chen et al. (2010)

recently demonstrated the impacts of inappropriate modeling of LCVs (Palardy & Vermunt, 2010) or of the nested data structure (Chen et al., 2010) on the estimates of individual and group effects (i.e., slopes and intercepts) as well as their predictors. As stated before, growth curve (and related) modeling methods have incredible potential for educational research as well as for decision making and policies that are based on evaluations, but for these methods to be both useful and used appropriately, the impacts of LCVs and hierarchical data within the growth curve modeling framework need to be fully investigated, particularly at the level of individual estimates (i.e., a parameter for each case) rather than at the effect level (e.g., overall group effect).

I build on the results of these three key studies (Chen et al., 2010; Muthén & Asparouhov, 2009; Palardy & Vermunt, 2010; Yumoto, 2011); I incorporated the PeLP student type (Chudowsky et al., 2007; Lazarus et al., 2010) to estimate, and understand the magnitude of, bias in estimates at each stratum of my analysis. As described above, there are two issues in the identification of LCVs, namely the assignment of individuals to levels of these variables and the appropriate estimation of effects of interest in MLGMM:

1) Covariates affect the identification of LCVs; and,

2) The nested structure has a limited impact on the identification of LCVs but can influence estimation and interpretation of random effects.

Coupled with the potential importance of the MLGMM for education research and decision-making, the salience of the LCV described by Muthén and Asparouhov (2009) and the substantively important class of PeLP students indentified by Lazarus et al. (2010) and Chudowsky et al. (2007) in their analyses, this body of work motivated me to quantify these effects in my work.

2.4 Algebraic Descriptions of MLM

The following are algebraic representations of the evolution of the models from a basic MLM (used in the traditional VAM) to more complex models such as the MLGMM, which is the model I used for my work. MLGMMs have multilevel capacities which can be used to incorporate the cluster level effect (also known as the between effect or teacher/school effect) and also have the capability to identify types or subpopulations of students in the data. VAM is a special case of the MLGMM; it is equivalent to an MLGMM where there is only one class, i.e., there is no mixture because everyone is assumed to be in the same class. When MLGMM is used instead of VAM, because it does include LC estimation, its use does not require an assumption that all students are in the same class or population.

2.4.1 Algebraic Description of Longitudinal Multilevel Model

The formulation for a two-level unconditional MLM (i.e., without covariates or explanatory variables) is:

Level-1

$$Y_{ti} = \pi_{0i} + \pi_{1i} T_{1ti} + e_{tij}, e_{ti} \sim N(0, \sigma^2) \quad (1)$$

Level-2

$$\pi_{0i} = \beta_{00} + \beta_{01} X_i + r_{0i} \quad (2)$$

$$\pi_{1i} = \beta_{10} + \beta_{11} X_i + r_{1i} \quad (3)$$

$$r_{ij} \sim N(0, \Theta_r) \quad (4)$$

where Y_{ti} is a response variable of the predicted student score at time *t*, T_{ti} are covariates at time *t* for individual *i*, *t* is a time or measurement occasion (e.g. 0, 1, 2, or 3), i is an individual, X_i is a time-invariant covariate which describes person characteristics (e.g. student SES) and e_{ti} is an error term at time *t* for individual *i* assumed to be normally distributed with homogenous variance across individuals. In the MLM formulation, repeated observations are nested within students where π_{0i} is an intercept or initial status for individual *i*, π_{1i} is a slope or growth rate for individual *i*, β_{00} is the mean initial status, β_{01} is the mean initial status difference between the defined student covariates X_i (e.g. student SES), β_{10} is the average growth rate, β_{11} is the mean difference in growth rate between the defined student covariates X_i , r_{0i} is the unique effect of individual *i* on mean initial status holding X_i constant, and r_{1i} is the unique effect of individual *i* on growth rate holding X_i constant. In addition, r_{0i} and r_{1i} are assumed to be random variables with zero means with variance-covariance matrix Θ_r representing the variability in the growth parameters remaining after controlling for X_i. This framework models student differences in the growth parameters based on some manifest student characteristics and it also assumes that the growth parameters, intercept and slope, are sufficient to capture individual growth trajectories.

2.4.2 Algebraic Description of Latent Growth Model

The formulation of an LGM is similar to that of an MLM. Equations 1 through 3 below are almost identical to Equations 5-7 for LGM. The main difference is that under the SEM framework time and student levels are subsumed under one single level, and they are both considered student observations (also known as the within-group part of the model) as opposed to the two different levels in MLM. LGM is expressed with the following four equations:

Within-group level measurement model

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{1ti} + e_{ti}, e_{ti} \sim N(0, \sigma^2) \quad (5)$$

Within-group level structural model for the intercepts and slopes

o Intercepts

$$\pi_{0i} = \beta_{00} + \sum_{m=1}^{M} \beta_{0m} X_{mi} + r_{0i} \quad (6)$$

o Slopes

$$\pi_{1i} = \beta_{10} + \sum_{m=1}^{M} \beta_{1m} X_{mi} + r_{1i} \quad (7)$$

$$r_{ij} \sim N(\mathbf{0}, \mathbf{\Theta}_{\mathbf{r}}) \quad (8)$$

In LGM, two growth factors, representing intercept and slope, completely capture individual growth trajectories as did the MLM.

2.4.3 Algebraic Description of Multilevel Latent Growth Model

The formulation of MLLGM is identical to that of an LGM with the addition of a level of analysis (the cluster level *j*) resulting in two levels of equations for MLLGM within the SEM framework (i.e., student observations for within-group and school observations for between-group). I define the term *cluster* as the grouping unit at level-2 in which students are nested. For my work, the cluster (also known as between or group) level represents the school level, more

specifically students' observations nested within schools. The formulation for the within-group part of the model is similar to LGM, except that instead of assuming r_{0i} and r_{1i} are random variables with zero means after controlling for X_i , I assume that r_{0i} and r_{1i} are random variables within a cluster (e.g. school) after controlling for X_i . In this framework, the outcomes for growth parameters depend on student characteristics within the same school and the parameters for student characteristic depend on school characteristics within the same school. For MLLGM Y_{tij} is the outcome at time *t* for individual *i* in school *j*, π_{0ij} is the initial status for individual *ij*, that is, the expected outcome for that individual at time zero, π_{1ij} is the growth rate for student *ij* during the academic year, and the random effect $e_{tij}$is assumed to be normally distributed with a mean of zero and variance σ^2. Also, β_{00j} is the mean status in school j for an individual with student characteristic X_i, β_{0mj} is student characteristic X_i gap on initial status, β_{10j} is the growth rate for a student with student characteristic X_i in school *j*, β_{1mj} is the student characteristic X_i gap on the academic year learning rate in school *j*, and random effects r_{0ij} and r_{i1j} are assumed to be normally distributed with a mean of zero and variance-covariance matrix Θ_r within the same cluster.

Level-2 models school characteristics where W_{nj} is a school characteristic used as a predictor for the school effect, γ_{000} and γ_{100} are the intercept terms in the school level model, γ_{00n} and γ_{10n} are the corresponding between level coefficients that represent the direction and strength of association between school characteristics W_{nj} and β_{mij} , and u_{0j} and u_{1j} are school level random effects that represent the deviation of school *j*' s coefficient , β_{mij}, from its predicted value based on the school level model. Level-2 random variables u_{0j} and $u_{1j}$are assumed to be random variables with zero means with variance-covariance matrix Θ_u. For the random effects, β_{00j} is the mean initial status in school *j*, β_{0mj} is the gap for covariates X_i on

initial status in school j controlling for school level covariate W_{nj}, β_{10j} is the growth rate for covariates X_i in school j, and β_{1mj} is the growth rate gap for covariates X_i in school j controlling for school level covariate W_{nj}. An MLLGM can be formulated with the following equations:

Within-group level measurement model

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} a_{1tij} + e_{tij}, e_{tij} \sim N(0, \sigma^2) \quad (9)$$

Within-group level structural model for the intercepts and slopes

o Intercepts

$$\pi_{0ij} = \beta_{00j} + \sum_{m=1}^{M} \beta_{0mj} X_{mi} + r_{0ij} \quad (10)$$

o Slopes

$$\pi_{1ij} = \beta_{10j} + \sum_{m=1}^{M} \beta_{1mj} X_{mi} + r_{1ij} \quad (11)$$

$$r_{ij} \sim N(\mathbf{0}, \mathbf{\Theta}_{\mathbf{r}}) \quad (12)$$

Between-group level model

o Intercepts

$$\beta_{0ij} = \gamma_{000} + \sum_{n=1}^{N} \gamma_{00n} W_{nj} + u_{0j} \quad (13)$$

o Slopes

$$\beta_{10j} = \gamma_{100} + \sum_{n=1}^{N} \gamma_{10n} W_{nj} + u_{1j} \quad (14)$$

$$u_{ij} \sim N(\mathbf{0}, \boldsymbol{\Theta}_{\mathbf{u}}) \quad (15)$$

2.4.4 Algebraic Description of Growth Mixture Model

Formulation of GMM models is achieved by adding a latent variable C_{ki} from LGM Equations 5 through 8 above. In addition, the GMM is a special case of MLGMM where no between-group models are included, resulting in the following specifications:

Individual level measurement model

$$Y_{ti} = \pi_{0i} + \pi_{1i} a_{1ti} + e_{ti}, e_{ti} \sim N(0, \sigma^2) \quad (16)$$

o Individual level structural model for the:

o Intercepts

$$\pi_{0i} = \sum_{k=1}^{K} \beta_{00k} c_{ki} + \sum_{m=1}^{M} \beta_{01k} X_{mi} + r_{0i} \quad (17)$$

o Slopes

$$\pi_{1i} = \sum_{k=1}^{K} \beta_{10k} c_{ki} + \sum_{m=1}^{M} \beta_{11k} X_{mi} + r_{1i} \quad (18)$$

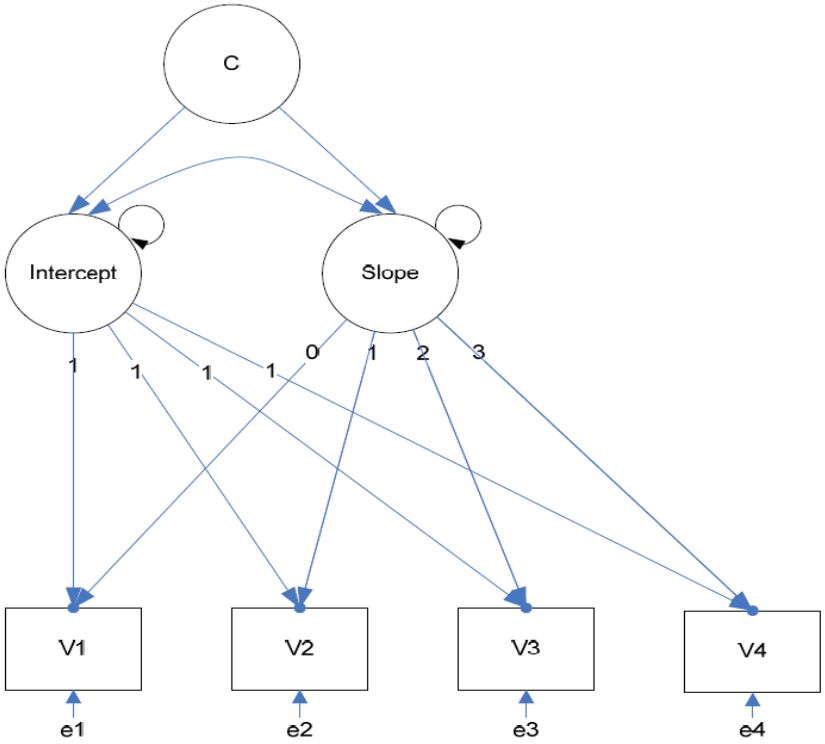
$$r_i \sim N(\mathbf{0}, \mathbf{\Theta}_{\mathbf{r}}) \quad (19)$$

Model for the latent class variables

$$\text{logit}[P(c_{ki} = 1)] = \delta_{0k} + \sum_{m=1}^{M} \delta_{mk} X_{mi} \quad (20)$$

It is assumed that individual growth factors are sufficient to estimate the effects of interest in the data, which can be seen in equations 17 through 20, by the exclusion of cluster *j*, notated in equations above 13-15, does not appear in any model. Figure 1 is a graphic representation of GMM, which is the same as the within-subject part of Figure 2 showing MLGMM

Figure 1. Graphical Description of GMM



2.4.5 Algebraic Description of Multilevel Growth Mixture Model (MLGMM)

The GMM (Muthén 2001; Muthén 2004) is a mixture extension of the LGCM or the LGM, and MLGMM is a multilevel extension of GMM (see Figure 3). The formulation of MLGMM builds on LGM by adding another level of analysis (between level), resulting in a MLLGM, and by adding LCVs *C* and *D* (as in GMM) from Figure 2 and any connections from/to these LCVs. The amalgamation of multilevel analysis and mixture modeling results in two levels of equations for MLGMM (i.e., one level for within-group and one level for between-group). . I include a description of MLGMM in this section to show how GMM and LGM are special cases of MLGMM.

The formulation of MLGMM has two parts: the within-group (i.e., level-1) and between-group (i.e., level-2) models. This formulation includes both within-group level and between-group level latent class variables. I focused on the estimates from the between-level slope, but the entire formulation is presented below for context (Yumoto, 2011).

Level-1

Within-group level measurement model

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} a_{1tij} + e_{tij}, e_{tij} \sim N(0, \sigma^2) \quad (21)$$

where π_{0ij} is an intercept for individual *i* in cluster *j*, π_{1ij} is a slope for individual *i* in cluster *j*, a_{1tij} is the set of covariates at time *t* for individual *i* in cluster *j*, and e_{tij} is an error term at time *t* for individual *i* in cluster *j* .

Within-group level structural model for the intercepts and slopes (Level-1)

o Intercepts

$$\pi_{0ij} = \sum_{k=1}^{K} \beta_{00k} c_{kij} + \sum_{m=1}^{M} \beta_{0jk} X_{mij} + r_{0ij} \quad (22)$$

o Slopes

$$\pi_{1ij} = \sum_{k=1}^{K} \beta_{10k} c_{kij} + \sum_{m=1}^{M} \beta_{1jk} X_{mij} + r_{1ij} \quad (23)$$

$$r_{ij} \sim N(\mathbf{0}, \mathbf{\Theta}_{\mathbf{r}}) \quad (24)$$

Model for subjects' latent class memberships, given their covariate

$$\text{logit}[P(c_{kij} = 1)] = \delta_{0k} + \sum_{m=1}^{M} \delta_{mk} X_{mij} \quad (25)$$

Level-2

Between-group level model

o Intercepts

$$\beta_{0ij} = \sum_{l=1}^{L} \gamma_{000l} d_{lj} + \sum_{n=1}^{N} \gamma_{00n} W_{nj} + u_{0j} \quad (26)$$

o Slopes

$$\beta_{10j} = \sum_{l=1}^{L} \gamma_{100l} d_{lj} + \sum_{n=1}^{N} \gamma_{10n} W_{nj} + u_{1j} \quad (27)$$

$$u_{ij} \sim N(\mathbf{0}, \boldsymbol{\Theta}_{\mathbf{u}}) \quad (28)$$

Model for Between-group for the latent class variable and class membership.

$$\text{logit}[P(d_{lj} = 1)] = \upsilon_{0k} + \sum_{n=1}^{N} \upsilon_{nk} W_{nj} \qquad (29)$$

where

t : time point

i : individual

j: group/cluster

a_{1tij}: individual level, time related variable

$X\ ij$: within-group level covariate

W_j: between-group/cluster level covariate

Equations 21 through 25 show the within-group (student) level models and Equations 26 through 29 show the between-group (school) level models. In equation 21, $Y_{tij}$is the observed individual outcome at time/occasion t for individual i within a group/cluster j (e.g. school), π_{0ij} is the expected value of Y_{tij} for this individual when t=0 (i.e. initial status), π_{1ij} is the expected slope/growth on the outcome for this individual (i.e. growth rate), a_{1tij}measures the time/occasions for this individual and e_{tij} is the residual/error term associated with this model for this individual. It is possible to include more time/occasion variables to model other growth effects (e.g. quadratic effect) in addition to the linear growth effect shown here. Equations 22 through 24 show the within-group (model of student differences) level model or the repeated measure for intercepts and slopes and Equation 25 shows the model for subjects' latent class memberships, given their covariates. Within-class intercepts and slopes are expressed with K-1

factors, *m* covariates X_{mij}, *k* latent classes C_{kij} , and random effects r_{0ij}. C_{kij} is equal to one when an individual *i* in cluster *j* belongs to the latent class *k* and otherwise zero where *k* = 1, 2,3,….,*K* and *K* is the total number of within-group latent classes, meanwhile $\beta_{0jk}$and β_{1jk} are the mean intercept and slope value for within-group class *k*. Equation 25 represents a multinomial logistic regression to describe the likelihood of membership in each of the latent class variable's levels, associated with predictors where *k*=1 is the reference class level.

Between-group (model of school differences) level equations 26 through 29 are almost identical to within-level equations from 22 to 25. Within-group heterogeneity in intercepts and slopes are regressed on three factors: between-group covariates, W_{nj} , between-group latent class variable, d_{lj}, and random effects (u_{0j} and u_{1j}) where *d* is the between-group latent class variable with *l* levels, and *L* is the total number of between-group latent classes (l = 1, 2, 3,….,*L*). $d_{lj}$is one when a cluster *j* belongs to the LC l and otherwise zero. γ_{0l} and γ_{1l} are the mean intercept and slope value for between group latent class variable level l. Equation 29 represents a multinomial logistic regression describing the likelihood of class membership associated with predictors where *k*=1 is the reference class level. The errors/residuals in each of the within-level measurement models, within-level structural/repeated measure models, and between-group models, are all assumed to be normal, independent across levels (e.g., between level-1 and level-2), and uncorrelated with the covariates. Figure 2 is a graphical representation of unconditional MLGMM based on the Muthén and Muthén (1993-2015) representation for the *MPlus* software.

Figure 2. Graphical Description of MLGMM

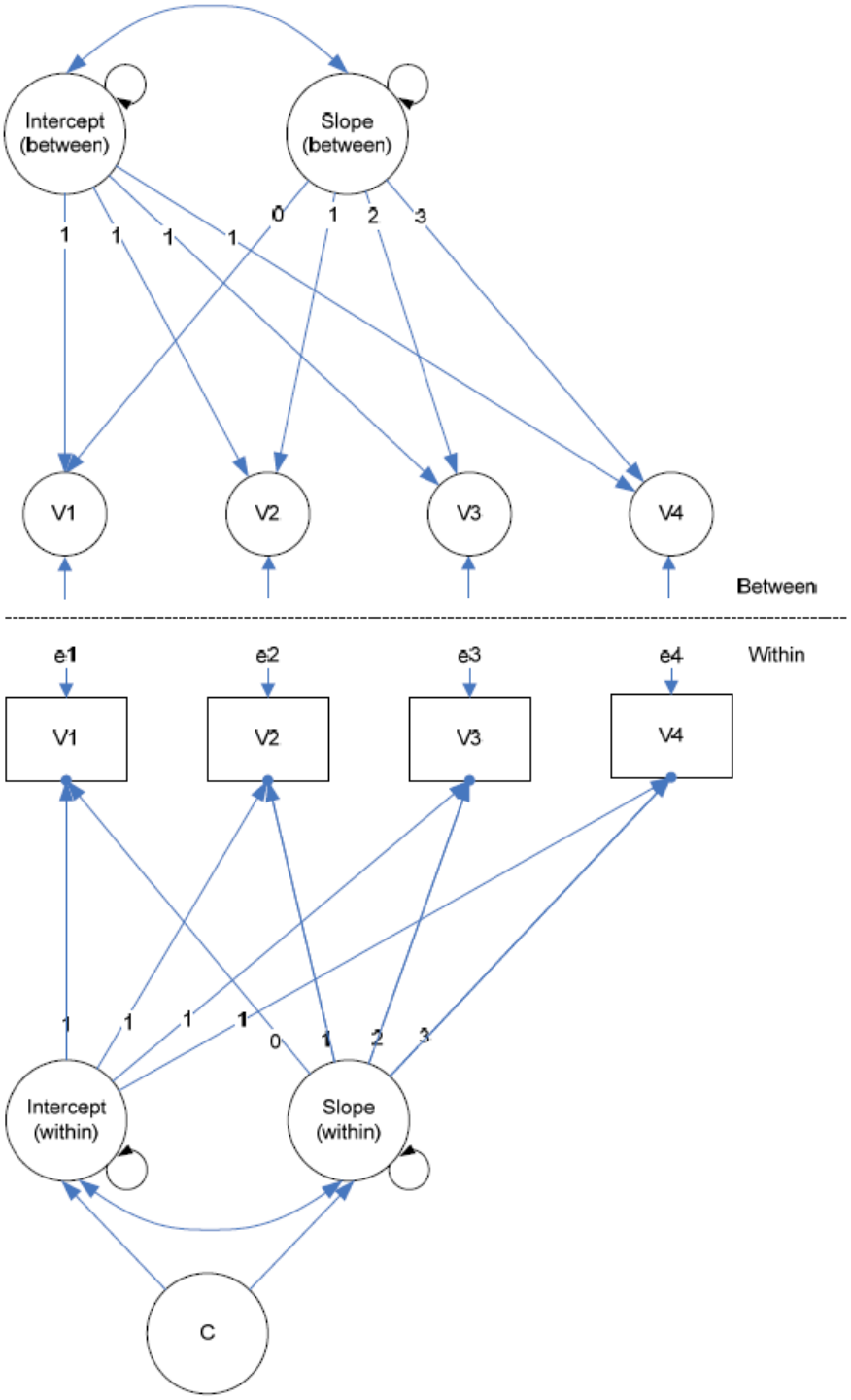
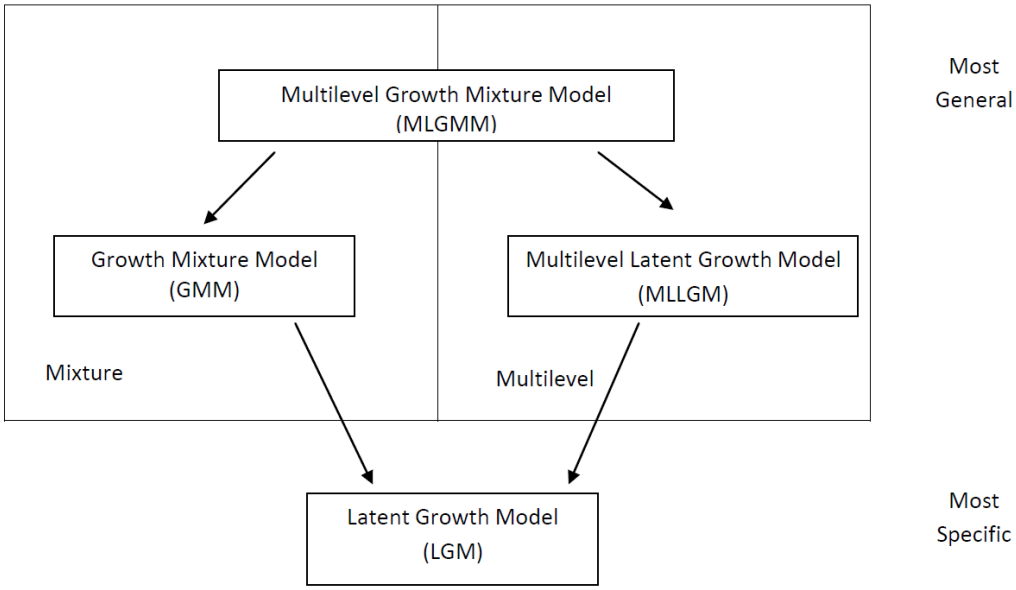


Figure 3. Advanced MLMs



2.5 Purpose of this Study

My primary focus in this research is to investigate the impact of ignoring the different types of performing school groups on the evaluation of school's effect on students' gain in test scores, focusing on the classification in the estimated school value-added scores. Most of the literature is focused on teachers at the cluster level but I will focus on schools at the cluster level since the models used are identical. I analyzed data from 3,360 students from grades 3rd to 6th motivated by the situation where school effect on student performance must be measured to evaluate the school's quality. In this situation, I will deduce the different types of students in the school based on their growth trajectory profiles in the past four years. I hypothesized that I would find four group profiles (HP, S, LP and PLP) in terms of the skills they are being taught, represented by both gains on standardized test scores (slope) and initial achievement level

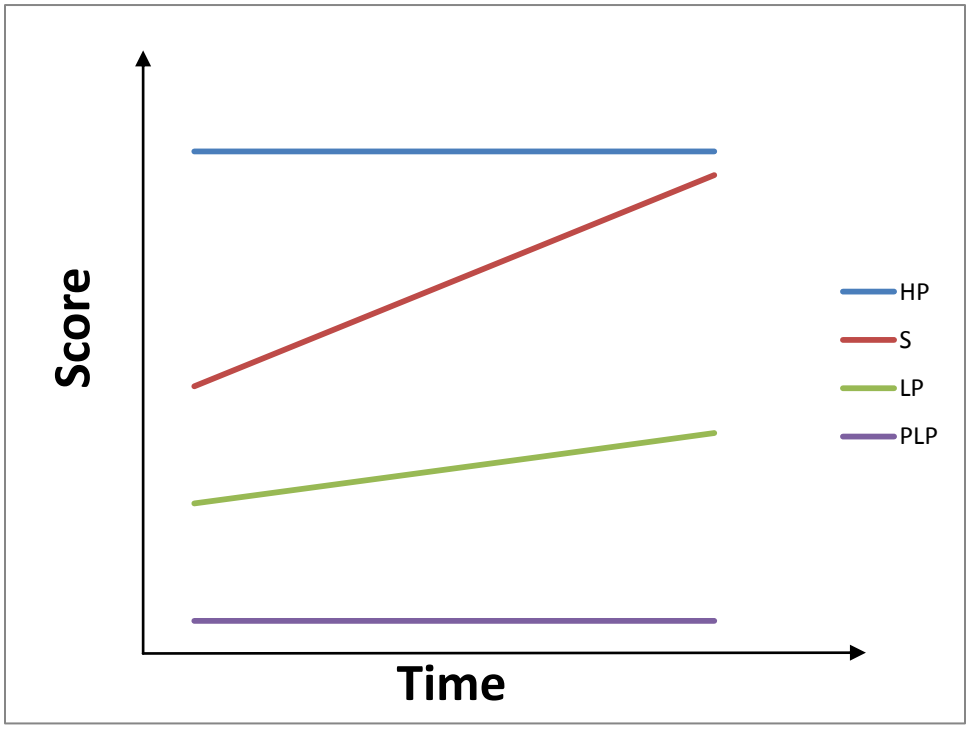
(intercept). Figure 4 shows a graphical representation of the expected students' growth profiles in each of these four groups (the actual slopes for particular students vary around these four lines). Each school has different proportions of students within each growth profile, which I represented in order to determine whether unknown, or ignored, heterogeneity in student type (based on proportion of students with each growth trajectory profile) inconsistent with the VAM assumption that all students get the same effect from a given school affected VAM-based estimates. There are two variables at the student level which are commonly integrated in VAMs: student SES and whether the student is LEP.

In addition, there is one school level covariate: school SES. With this example, imagine that the proportion of students in the four growth groups is different among different types of schools' SES levels, resulting in the overall achievement level of that school. I posit that only students in the S group receive any benefit of instruction from schools –that is, the school's effect is not zero and positive for students in these two groups.

In this scenario, it is very difficult for schools with majority PLP students to obtain a high value added score as compared to schools with majority HP achieving students – even if they have added identical value compared with schools in high achieving classes – because the expected average school effect is attenuated by the group of students who are not responsive to any instruction. In other words, schools are penalized, in terms of the estimation of their effectiveness, by the kinds of students they have in the classroom under the assumption that all students receive the same benefit from the instruction. Thus, a sound accountability system for schools evaluation cannot be established without accounting for the growth profile (type) of students.

I investigate the classification change in the school value-added score when student type is un-modeled, that is, under the VAM assumption that the students receive a homogeneous (randomly, not systematically, varying) effect from the school, by manipulating conditions identified and described more extensively in Chapter 3.

Figure 4. Graphical Description of Growth Profiles



CHAPTER III

METHODOLOGY

In this section, I describe this study cohort, the characteristics of the empirical study including manifest variables (at Level-1 and at Level-2) with MLM and MLGMM frameworks. Finally, I also specify the corresponding hypotheses.

3.1 Cohort Description

My data comes from a mid size urban district in North Carolina. To hold constant other confounding sources in my analysis, my study population was restricted to those students who were enrolled in the school district for the four consecutive years of my work (grades 3 through 6). This population was further restricted to students who attended both the same elementary school and middle school. The resulting study population consists of 3,360 students, who attended 41 different schools. The demographic composition is as follows: half females (50%) and half males (50%); their ethnicity was mainly Caucasian (41%), followed by African Americans (28%) and Hispanics (24%). The majority (58%) of students had an FRL-58%) identification and about a quarter (24%) were identified as LEP. Most LEP students were of Hispanic descent (89%) and are in FRL (92%). This cohort took the state's standardized assessment in reading over 4 years: in 2011 as third graders, 2012 as fourth graders, in 2013 as fifth graders and in 2014 as sixth graders. Table 1 shows the summary statistics that characterize this cohort of students and schools in the state.

Table 1. Grades 3 through 6 Cohort Sample Description

Description	Reading Cohort
Students	3360
Schools	41
Female	50%
Black	28%
Hispanic	24%
White	41%
Free and reduced price lunch	58%
Students with disability	11%
English language learner at Time 1	24%
Unstandardized test score M (SD) Time 1	339.61 (12.04)
Unstandardized test score M (SD) Time 2	345.45 (10.61)
Unstandardized test score M (SD) Time 3	449.05 (10.16)
Unstandardized test score M (SD) Time 4	451.32 (11.17)

Note: The students taking the reading test was in Grade 3 as of 2011 and in Grade 6 as of 2014

The students in my study population were somewhat more likely to be non-White, English language learners, and eligible for free and reduced lunch services than those students excluded from analysis. The students and schools selected for analysis also tended to have slightly lower average test scores. In addition, most African American students were in FRL status (80%) as were most Hispanic students (93%); fewer White students were so identified (23%). Since I also focused on the LEP population, the LEP students are described in more detail. Students' LEP designation status (which is determined by their yearly ACCESS score) changes with every year based on whether the students achieve a score of Level 4 (LEP status) or Level 5 (non-LEP status) in several domains. About 24% of the study population were identified as LEP in the first year (Time 1); 73% of these students retained their LEP status in the second year (Time 2), 54% retained their LEP status as of the third year (Time 3), but only 13% were still considered LEP by the fourth year (Time 4). Since I also focus on schools' SES (which I define as the percent of FRL students within a school), I describe the number of schools with a given SES in more detail. The district has 19 schools with more than 87% of their students (1311 students) in FRL, 8 schools with between 50%-87% of their students (828 students) in FRL and 14 schools with fewer than 50% of their students (1221 students) in FRL.

In 2013, during the time I conducted my study, changes were implemented to the North Carolina state curriculum to fulfill the new mandated standards of "Common Core." In addition to the curriculum changes, the End-of-Grade (EOG) test assessment scales (both reading and math) were modified from 302-367 to 406-462 for grade 3, from 313-370 to 412-468 for grade 4, from 319-375 to 418-473 for grade 5; and from 322-377 to 416-478 for grade 6. As a result of the curriculum modifications, the new reading test form dramatically impacted the percent of students who were deemed proficient for school year 2013. The percent of students who were not proficient are presented in Table 2. Changing the curriculum for the reading assessment in 2013

(Time 3, students were in Grade 5) increased the percent of students being identified as not proficient from 38% to 62%. The apparent decline in students' scores created a public protest and the test scores for 2014 (Time 4) were rescaled yet again, resulting in a decrease of 15% (62% to 47%) in the proportion of students being identified as not proficient for this particular population. Since the scales of the scores were changed twice (at Time 3 and Time 4), for the relevant portion of my analysis (i.e. school effect on student achievement) I used test scores standardized to have an M of zero and SD of one within each grade (Ballou and Springer, 2011). The several scale changes of the EOGs during the study period may be a potential limitation of my results. However, all studies have limitations, particularly in educational settings where curriculum and, consequently, scales are constantly changing. Scales for the EOGs remain constant for no more than five years (but three years is the norm). Standardizing the scores should be sufficient to not affect the results. The outcome variables (Reading EOG scores for the four year period) are continuous and because I am interested in the growth of students given a certain initial status, what matters is that the range of the scores remains the same and that the hierarchy of the student scores are also maintained. I believe I have met these conditions.

Table 2. Percent of Students Not-Proficient from 2011-2014

Percent Proficient	Grade 1	Grade 2	Grade 3	Grade 4
	Time 1	Time 2	Time 3	Time 4
Ntotal	3360	3296	3309	3360
% Not Proficient	41%	38%	62%	47%

Table 3 shows the summary statistics for LEP and non-LEP students on standardized scores for the resulting grade-specific M and SD estimates. The non-LEP estimates include students who were never identified as LEP for the four years of the study. The LEP category includes students identified as LEP at time 1 of the study.

Table 3. Summary Statistics Standardized Reading Scores

	Non-LEP		LEP	
Grade	M	SD	M	SD
3	0.258	0.943	-0.564	0.873
4	0.256	0.943	-0.537	0.868
5	0.272	0.924	-0.537	0.881
6	0.266	0.930	-0.534	0.882

The significant association (i.e. SES and minority ethnic groups) in commonly used manifest variables (e.g. student SES, ethnicity and LEP) in VAM models confounds the cause-effect estimation of growth parameters if modeling is based only on these manifest variables. The model may falsely indicate a strong weight for the ethnicity parameter when indeed ethnicity may

just be a distal predictor of student performance. When manifest variables have a strong association with student deficiency, whatever form this deficiency might take (i.e. extreme poverty, parents illiteracy, lack supportive network or others), the model (which does not specify relevant factors at level-1) may falsely indicate significant effects of growth (positive or negative) to the school when they may actually pertain to characteristics of the student.

3.2 Characteristics of my Empirical Study

I propose that there are two distinct sources of bias in estimated teacher/school effects that have not been addressed in the existing literature. These sources (which directly influence test scores) include: 1) the resources students received (e.g. extra tutoring), which are not included in the model; and, 2) the bias that is due to the systematic association of certain schools with students who are low performers (due to very low SES). Peisner-Feinberg (2015) and Garcia (2015) both discuss this student-selection-based bias. The first source of bias is a form of omitted variables bias that exists even if the students are randomly assigned to schools and the second source will influence the model results as long as individual schools are correlated with certain student characteristics (e.g., student SES). My main objective is to estimate school effects without conflating school effects and the effects of student self-selection to schools and treatments that were not controlled for in the model (e.g. extra tutoring). I will address the first source of bias at level-1 of the model through the introduction of latent classes and the second source at level-2 of the model by introducing a school level covariate: school SES. I use the EOG reading scores, which effectively require two levels of analysis: Level-1 is the longitudinal growth within each student in addition to the student level characteristics (i.e., LEP status, student SES and latent class) and Level-2 contains the school level characteristics (e.g., school SES). I used *Mplus 7* (Muthén & Muthén, 2016) to analyze the traditional longitudinal MLMs and the MLGMMs. Other statistical analysis such as the t-tests I used to assess value-added rankings

between schools I obtained using SAS (9.3, SAS Inc., Cary, NC). Since I used M*plus* (Asporouhov, 2009) to analyze my data, I only specify two levels of analysis. M*plus* has the capability to analyze time and student level data in one step but SAS offered the option of three levels of analysis (i.e. time, student and school) for this same study(e.g., repeated tests scores as level-1, repeated scores nested within students as level-2 and students nested within schools as level-3).

3.3 Characteristics of Individual (Level-1) Data

In this part of my analysis I analyze student observations (repeated observations and student characteristics). I also analyze the covariate estimates of the two model frameworks to assess how much the interpretation of these key student level covariates changes with the introduction of the latent class at level-1. I expect that by introducing a latent class at level-1 the student estimates and school estimates will be more attenuated (in magnitude) compared with the student estimates and school estimates obtainable from a traditional MLM (e.g., being a LEP or low SES student will have less of a negative impact on student growth when the latent class is specified in the model). The level-1 manifest variables I introduce into the model are student SES and LEP status. The student SES variable is discrete with two levels: High (the student did not receive FRL) or Low (the student received FRL). The LEP variable is also discrete with two levels: Yes (if the student was assigned LEP status at Time 1) and No (if the student was never assigned LEP status). Although LEP identification changes over time, I expect that the latent class will better describe the variability of this population.

3.3.1 Latent Classes Identification

I argue that a potentially large source of heterogeneity still resides in the variation of the regression coefficients due to the presence of a mixture of distributions. This interdependence

between groups of students sharing similar but unobserved background characteristics is captured by my level-1 latent classes (student level). The most reasonable number of latent classes to extract I determined from the data. I used a multinomial model (Equation 25) with level-1 time (intercept and growth rate at level-1) variables as predictors. I determined the latent classes only from the random intercepts (initial status) and slopes (growth rates) of the time level variables (e.g., unconditional model). Specifically, the time level-1 intercept represents the individual starting point in the reading EOG scores and the slope represents the growth rate of that individual from one year to the next. I hypothesize that four class levels represent different growth trajectories in individuals (i.e., at level-1); in other words, these four class levels capture the level-1 heterogeneity in my data. I believe that my use of the growth profiles (i.e. initial status and growth rates) of these four groups expands on the work of Yumoto (2011) and that of Chen et al. (2010), which in turn is based on Nylund et al. (2007). These researchers hypothesized two latent profiles trajectories, namely one with steeper slope with a higher intercept and one with shallower slope with lower intercept. The scenario described by these researchers assumes that there is a group of students (who they called 'fast growing') who start at a high level and also have a high growth rate (Mean 2.5, Slope 0.6 with standardized scores with zero M and SD of one) and that a second group of students (who they call 'slow growing') exists who start at a lower level and also have a low growth rate (Mean 1, Slope 0.1 with standardized scores with zero mean and SD of 1). However, other scholars have found that students who are in the upper level of the scale (above one SD from the mean) in reading tests grow at a much slower rate in comparison with students in the lower level (between 0 and 1 SD from the mean) of the scale (Braun, 2005; Ballou and Springer, 2011). Thus, I expand the criteria to four intercept levels (one for each growth profile) and four slope levels (two with no growth and two with strong and low slopes). I use this scenario to depict more accurately students' growth trajectory, in line with prior research

in this area (Jennings, 2005; Krieg, 2008; 2011; Neal and Schanzenbach, 2007; Lauen & Gaddis, 2010; Dee and Jacob, 2011 and Loveless, 2008). Table 4 shows the hypothesized settings representing each growth profile (HP, S, LP and PLP) included within every MLGMM model. I expected that the HP group would have a high intercept mean with no growth, the S group would have the second highest intercept with a strong growth rate, LPs would have similar intercept as the S group but with low growth rate while the PLPs would have the lowest intercept mean and no growth. I do not assume that each of these growth profiles will contain an equal number of students.

Table 4. Growth Parameter Settings for the Four Latent Classes

Parameters	HP	S	LP	PLP
Intercept Mean	High	Low	Low	Very Low
Slope Mean	Zero	Strong	Low	Zero

3.3.2 Level-1 Estimation

I begin my initial analysis with level-1 estimation. Here, I focus on the student level variables as well as the differences in estimates of contribution of student growth (i.e. intercept and slope) when the data is modeled with a traditional MLLGM versus a MLGMM. Since *Mplus* analyzes these two levels in a single step, my Level-1 data is comprised of time (or longitudinal) and student level data (e.g. student SES and LEP). The time level (t=0, 1, 2, 3) is defined by four repeated EOG reading scores (grade 3 through grade 6). I place the reading scores in a standardized scale with an M of zero and SD of one to be able to measure student development over time. I specify the first five models using a traditional MLLGM framework. My initial model specification has only the growth factors (i.e., random intercept and slope) without any

manifest covariates (at level-1). My second model specification has one fixed manifest variable at the student level (student SES as previously defined using FRL). My third model specification has one fixed manifest variable (student SES) and one time varying manifest variable (student LEP identification). My fourth model specification adds the second level of analysis (school level) to the third model specified. My final model specification (full model) adds the level-2 covariate school SES to the fourth model specification.

I specify my next five models using a MLGMM framework. For the modeling process with MLGMM, I follow the steps described with the MLLGM framework except that in my initial model specification I specify my LCs in addition to specifying the growth factors (which I retain in the remaining four model specifications). My goal in this modeling process is to discern how much the interpretation of the student level covariates on student growth changes based when the analysis moves from an underspecified model to a model that is more complete with relevant variables (as I described in Chapter 2). For clarification, in the next section I present the algebraic representation of the complete model to be tested and I specify each of its components in detail.

3.4 Data Study Framework

To address my research questions, I specify a VAM that includes student characteristics at level-1 and within-school characteristics at level-2. I specify two frameworks that model the test-score outcomes : one framework without LCs at level-1 (a traditional MLLGM), the other framework with LCs at level-1 (an MLGMM). In a modeling context, the influence of omitted variables bias is resolvable by adding latent classes at level-1 and the student selection bias can be addressed by comparing schools with a similar student composition (based on level-2 school SES). By conditioning on LCs, I am able to obtain consistent estimates of the school effects as long as there is no selection of students to schools. Based on the background I provided in

Chapter 2, the following Equations (30-54) show the full models that I used to fit the data. Equations 30 through 40 represent a traditional MLLGM and equations 41 through 54 represent the MLGMM. Both model frameworks contain two level of analysis: the within-group (i.e., level-1 or student level) and between group (i.e., level-2 or school/teacher level). The level-1 indicators contain the observed outcome (Y_{tij}) for the individual *i* in cluster *j* at a given point in time, the growth parameters (π_{oij} and π_{1ij}), the covariate time (*t* =0, 1, 2, 3) and the unexplained random variation (e_{tij}) of individual *i* in cluster *j* at a given point in time (equation 30). I assume that the random error has a normal distribution with a M of zero and SD of 1 (equation 30).

The two growth parameters for both model frameworks contain the intercept or initial status for the individual *i* in cluster *j* represented by π_{oij} (equations 31 and 41) and the growth rate for individual *i* in cluster *j* represented by π_{1ij} (equation 32 and 42). In equations (31, 32) and (41, 42) the two growth parameters are outcome variables and are nested within students. Both growth parameters have predictors or covariates that contribute to the estimation of the initial status and growth rate of each individual. I express the within-class growth parameters (intercepts and slopes) using four components: two manifest covariates (i.e., student SES and LEP), one latent class (student growth profile) and random error (r_{0ij} and r_{1ij}). The random errors(r_{0ij} and r_{1ij}) are considered to be multivariate normal with a mean of zero and variance covariance matrix $\begin{bmatrix} \tau_{\pi_{00}} & \tau_{\pi_{01}} \\ \tau_{\pi_{10}} & \tau_{\pi_{11}} \end{bmatrix}$ (equation 33).

The main difference between the model frameworks (MLLGM and MLGMM) is the addition of the LC ($Class_{ij}$) at level-1 in the MLGMM (equation 41 and 42). The LCs represent the conditional student growth profile, for individual *i* in cluster *j*. Equations 37 through 39 represent the estimate of school effect when LCs are not specified in the model and student

trajectory performance profiles are not included in the model; equations 49 through 54 represent the estimate of school effect when , the estimate of school effect is given by equations 37 through 39 and when latent classes LCs are specified. The cluster level (level-2) is determined by school and school SES (defined using the percent of students receiving FRL for a given school) is the level-2 covariate. The school SES is inversely correlated with the percent of students receiving FRL: the greater the percentage, the lower the school SES, and vice versa. The two model frameworks I describe below utilize all the variables available : two manifest covariates (student SES and LEP) and one latent variable indicating class membership for level-1 and school SES for level-2.

In addition to the full model I describe below, I fit several other models with different combinations of variables mentioned above as part of my model building process. First, as a baseline model, I specified the level-1 unconditional model (i.e., only with growth factors) using the MLLGM framework. Second, I included student SES as the sole covariate predicting students' initial status and growth rate. Third, I added the variable LEP to the previous model as a second covariate to improve my specification of the model. Fourth, I add the school level to the analysis and finally I specify school SES at level-2. I performed this modeling process twice (five models without LC specified at level-1 and another set of five models when LC is specified at level-1) to understand how the interpretation of school value-added effects change when my models account for student (level-1) and school variability (level-2). In the below frameworks I only describe the full final model versions to keep the description of the two models simple.

The first framework, conventional two-level regression (MLLGM), is explained below,

Level-1

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}\, time_{1tij} + e_{tij}\,, e_{tij} \sim N(0,1) \tag{30}$$

$$\pi_{oij} = \beta_{00j} + \beta_{01j}\, Student\ SES_{ij} + \beta_{02j}\, LEP_{ij} + r_{0ij} \tag{31}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}\, Student\ SES_{ij} + \beta_{12j}\, LEP_{ij} + r_{1ij} \tag{32}$$

$$where\ \begin{bmatrix} r_{0ij} \\ r_{1ij} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi_{00}} & \tau_{\pi_{01}} \\ \tau_{\pi_{10}} & \tau_{\pi_{11}} \end{bmatrix}\right) \tag{33}$$

Level-2

Intercept

$$\beta_{00ij} = \gamma_{000} + \gamma_{001} SCH\_SES_j + u_{0j} \tag{34}$$

$$\beta_{01ij} = \gamma_{010} + \gamma_{011} SCH\_SES_j + u_{1j} \tag{35}$$

$$\beta_{02ij} = \gamma_{020} + \gamma_{021} SCH\_SES_j + u_{2j} \tag{36}$$

Slope

$$\beta_{10ij} = \gamma_{100} + \gamma_{101} SCH\_SES_j + u_{3j} \quad (37)$$

$$\beta_{11ij} = \gamma_{110} + \gamma_{111} SCH\_SES_j + u_{4j} \quad (38)$$

$$\beta_{12ij} = \gamma_{120} + \gamma_{121} SCH\_SES_j + u_{5j} \quad (39)$$

$$where\ u_{0j} \sim N\left(0, \tau_{\beta_{000}}\right) \quad (40)$$

The second framework, MLGMM, is explained below,

Level-1

$$\begin{aligned} \pi_{0ij} = {} & \beta_{00ij} + \beta_{01j}\, Student\ SES_{ij} + \beta_{02j}\, LEP_{ij} + \beta_{03k}\, Class_{ijk} \\ & + \beta_{04jk}\, StudentSES_{ij} * Class_{ijk} + \beta_{05jk}\, LEP_{ij} * Class_{ijk} \\ & + r_{oij} \end{aligned} \quad (41)$$

$$\begin{aligned} \pi_{1ij} = {} & \beta_{10ij} + \beta_{11j}\, Student\ SES_{ijk} + \beta_{12j}\, LEP_{ij} + \beta_{13k}\, Class_{ijk} \\ & + \beta_{14jk}\, StudentSES_{ij} * Class_{ijk} + \beta_{15jk}\, LEP_{ij} * Class_{ijk} \\ & + r_{1ij} \end{aligned} \quad (42)$$

Level-2

Intercept

$$\beta_{00ij} = \gamma_{010} + \gamma_{011} SCH\_SES_j + u_{oj} \quad (43)$$

$$\beta_{01ij} = \gamma_{010} + \gamma_{011} SCH\_SES_j + u_{1j} \quad (44)$$

$$\beta_{02ij} = \gamma_{020} + \gamma_{021} SCH\_SES_j + u_{2j} \quad (45)$$

$$\beta_{03ij} = \gamma_{030} + \gamma_{031} SCH\_SES_j + u_{3j} \quad (46)$$

$$\beta_{04ij} = \gamma_{040} + \gamma_{041} SCH\_SES_j + u_{4j} \quad (47)$$

$$\beta_{05ij} = \gamma_{050} + \gamma_{051} SCH\_SES_j + u_{5j} \quad (48)$$

Slope

$$\beta_{10ij} = \gamma_{100} + \gamma_{101} SCH\_SES_j + u_{6j} \quad (49)$$

$$\beta_{11ij} = \gamma_{110} + \gamma_{111} SCH\_SES_j + u_{7j} \quad (50)$$

$$\beta_{12ij} = \gamma_{120} + \gamma_{121} SCH\_SES_j + u_{8j} \quad (51)$$

$$\beta_{13ij} = \gamma_{130} + \gamma_{131} SCH\_SES_j + u_{9j} \quad (52)$$

$$\beta_{14ij} = \gamma_{140} + \gamma_{141} SCH\_SES_j + u_{10j} \quad (53)$$

$$\beta_{15ij} = \gamma_{150} + \gamma_{151} SCH\_SES_j + u_{11j} \quad (54)$$

The parameters β_{01j} and β_{02j} are the contributions from the covariates student SES and LEP (respectively) to the intercept (initial status) and the parameters β_{11j} and β_{12j} are the contributions from the covariates student SES and LEP to the slope (growth rate), for individual *i* in cluster *j* for both the conventional MLLGM and the MLGMM (equations 31, 32 and 41 and 42, respectively). The parameters β_{03j} and β_{13j} are the contributions from the LC to intercept and slope for individual *i* in cluster *j* for the MLGMM (equations 41 and 42, respectively). The parameters β_{04j} and β_{14j} are the contributions from the interaction term of covariate student SES and LC to the intercept and slope for individual *i* in cluster *j* for the MLGMM (equations 41 and 42, respectively). The parameters β_{05j} and β_{15j} are the contributions from the interaction term of covariate LEP and LC to the intercept and slope for individual *i* in cluster *j* for the MLGMM (equations 41 and 42, respectively).

Between-level parameters are described by equations 34 through 40 for the conventional MLLGM and equations 43 through 54 for the MLGMM. The components of the between-level parameters are: within-group heterogeneity in intercepts (e.g. β_{01j}), and slopes (e.g. β_{11j}), level-2 covariate (school SES) and random effects (u_{0j} through u_{5j} in equations 34 through 39 for MLLGM and u_{0j} through u_{11j} in equations 43 through 54 for MLGMM). The within-group intercepts and slopes are nested within schools and they are included in a model with no covariates at level-2 (unconditional at level-2) or with a between group covariate: school SES (conditional on school SES at level-2).

The parameters γ_{011} and γ_{111} are the effects for the between covariate school SES for both model frameworks on the level-1 covariate student SES which respectively affect level-1 student initial status and growth rate. The parameters γ_{021} and γ_{121} are the effects for the between covariate school SES for both model frameworks on the level-1 covariate LEP which respectively affect level-1 initial status and growth rate. The covariates γ_{031} and γ_{131} are the effects for the between covariate school SES for the MLGMM framework on the level-1 latent class which respectively affect level-1 initial status and growth rate. The covariates γ_{041} and γ_{141} are the effects for the between covariate school SES for the MLGMM framework on the level-1 interaction term of LC with student SES which respectively affect level-1 initial status and growth rate. The covariates γ_{051} and γ_{151} are the effects for covariate school SES for the MLGMM framework on the level-1 interaction term of LC with LEP which respectively affect level-1 initial status and growth rate. Note that for models without a covariate at level-2 (without school SES), the parameters of interest are: γ_{010} and γ_{110} (which affect the level-1 covariate student SES), γ_{020} and γ_{120} (which affect the level-1 covariate LEP), γ_{030} and γ_{130} (which affect the Level-1 LC), γ_{040} and γ_{140} (which affect the interaction term of the Level-1 covariate student SES with the LC), and γ_{050} and γ_{150} (which affect the interaction term of the Level-1 covariate LEP with the LC).

In the following section, I describe the two manifest (at -Level-1) variables in the models and I specify the corresponding hypotheses.

3.5 Student SES as Level-1 Covariate

I chose student SES to be specified in the model as a time-invariant covariate at the student level because it is a relevant variable and has a direct effect on student learning gains (Garcia, 2015). I fit my data using two different model frameworks. First, I model my data using

a traditional MLLGM to obtain the effect estimate of student SES on student growth parameters (initial status and growth rate) as my baseline model. Second, I model my data using an MLGMM to assess how much the estimates of student SES on student growth parameters differ (magnitude and direction) from the traditional model. In the MLGMM case, in addition to the manifest variable mentioned above, I add the LC (i.e., student growth performance profile) to this level as an additional student characteristic. I assume that this LC contains a proportion of the "truth" in statistical identification. The LC may account for a good portion of the variation in the random intercept and slope (i.e., initial point and growth rate) because there are several relevant omitted variables that also account for student achievement not specified in the model.

As I mentioned previously, student SES is insufficient to explain student growth due to the diversity students within the same SES (Garcia, 2015); for instance, although low SES students tend to be low performers, there are also some moderating mechanisms by which student SES indirectly affects growth through other external factors (e.g. parental engagement). Several other variables (e.g., resources students received) which support school achievement also interact with student SES (Garcia, 2015). Since schools do not have access to these other external factors, I present a methodological solution to try to account for the remaining systematic variability that can be captured with an LC, i.e., the students' performance profiles (Yumoto, 2011). As a result, in the model specification the manifest variable (student SES) and the latent class have a direct influence on the random intercept and random slope, but the manifest variable also has an indirect influence on the random intercept and random slope via the LC (MacKinnon, 2008; Muthen & Muthen, 2010). I expect that the mixture model results will challenge the interpretations that are obtained from the conventional MLLGM. I also expect that the gap between high and low SES students is smaller and probably not statistically significant when assessed using the MLGMM in comparison to the gap found with the traditional model because the parameters for high SES

($\beta_{01HighSES}$ and $\beta_{11HighSES}$) and low student SES ($\beta_{01LowSES}$ and $\beta_{11LowSES}$) will be more attenuated. More specifically the magnitude of the parameter for high SES students will be smaller and positive and the magnitude for the low SES students will be smaller and negative, resulting in the narrowing of the gap with MLGMM. As a result, MLGMM($\beta_{01highSES} - \beta_{01lowSES}$) < MLLGM($\boldsymbol{\beta_{01highSES} - \beta_{01lowSES}}$) and MLGMM($\beta_{11highSES} - \beta_{11lowSES}$) < MLLGM ($\beta_{11highSES} - \beta_{11lowSES}$).

3.6 Adding LEP as Level-1 Covariate

Since NCLB applies to all public schools, it is unclear what comparison group provides a relevant counterfactual; however, LEP students comprise one of the groups drawing attention because they are the fastest growing population in the country, comprising about 10% of students in the U.S public schools (Ruiz Soto, Hooker and Batalova, 2015). For this reason, I also specify LEP when assessing student level estimates of the traditional model from the MLGMM. The LEP variable is the only level-1 time variant covariate, with two response levels as previously defined. In addition, LEP students have received English as a Second Language (ESL) services for about 4 years in the district (the LEP students included in this study were not new to the district or neither they were newly identified). For this reason, given that LEP students had English instruction for some time, I expect some variability in their initial status and growth rate estimates. This variability in initial status and growth rate can sort LEPs into different hypothesized performance profiles described in section 3.3.1.

First, I add the LEP variable to the MLLGM with only student SES as a covariate, resulting in a model with a traditional framework with two manifest variables (student SES and LEP). Second, I model the data using a MLGMM with an LC added to the student level. I assume that this LC contains part of the "truth" in statistical identification and may ultimately

affect school level estimation. The LCs may account for some of the systematic variation in the random intercept and slope (i.e., initial point and growth rate) with some further systematic variation across student level units captured by the LEP variable plus random variation. I expect that the LCs will capture the remaining systematic variability of other moderators not specified in the model. I expect that the mixture model results will challenge the interpretations I obtain from the traditional model, since the traditional model assumes that student SES and LEP are enough to account for all the relevant variability in student achievement, whereas the MLGMM only partly attributes this relevant variability to LEP and student SES. There is still relevant variation in intercept and slopes even after controlling for LEP and student SES which I can capture using the growth performance profile of the student because the profile includes other (omitted) variables with an impact on student achievement.

From an intervention point of view, the traditional model and the mixture model may lead to different decisions on what to manipulate. For example, the traditional model may indicate that LEPs are performing significantly worse than non-LEPs and its results may suggest harsher punishments for schools with a disproportionate number of LEP students who are LPs and PLPs. However, the MLGMM (which controls for LC in addition to LEP) may indicate that HPs and S LEPs and non-LEPS perform similarly. However, among LPs and PLPs, LEPs may be performing worse, which would suggest that a more stringent criterion should be established to take them out of the LEP identification, since it would be detrimental to their learning development to remove prematurely the language support the school provides for these two subgroups of LEPs. The traditional model ignores the membership in a given class, which, if acknowledged, would make the issue a student level decision, thus decision makers would discern that a great part of the growth of a student depends on the student's profile. Introducing LC to the model permits for a more complete observation of students, thus decision makers

consequently would conclude a school level intervention (e.g. closing a school) might be less effective than a student level intervention (Schochet & Chiang, 2013)

I present the formulation of a conditional level-1 MLLGM in equations 31 and 32. I hypothesize that in this model the parameters β_{02j} and β_{12j} are moderately higher for students who belong to a non LEP status than they are for students who belong to LEP status when controlled for student SES. As a result, the parameter estimates $\beta_{02nonLEP} > \beta_{02LEP}$ and $\beta_{12nonLEP} > \beta_{12LEP}$. With regards to the student SES covariate, I hypothesize that the parameters β_{01j} and β_{11j} are moderately higher for high SES students than they are for low SES students when I control for student LEP. The estimates slightly change from being strongly higher to moderately higher when LEP is added in the model because LEP takes some of the variability. As a result, the parameter estimates $\beta_{01highSES} > \beta_{01lowSES}$ and $\beta_{11highSES} > \beta_{11lowSES}$. I expect the most dramatic change to occur when the LC is introduced to the model.

I present the formulation of a conditional level-1 MLGMM in equations 41 and 42. The formulation has two manifest covariates at the student level (student LEP and student SES) and one LC. I hypothesize that in the MLGMM the parameters β_{02j} and β_{12j} are mildly higher for students who belong to non LEP status than they are for students who belong to LEP status when controlling for student SES and LC. As a result, MLGMM $(\beta_{02nonLEP} - \beta_{02LEP}) <$ MLLGM $(\beta_{02nonLEP} - \beta_{02LEP})$ and MLGMM $(\beta_{12nonLEP} - \beta_{12LEP}) <$ MLLGM $(\beta_{12nonLEP} - \beta_{12LEP})$. With regards to student SES covariate, I hypothesize that in the MLGMM the parameters β_{01j} and β_{11j} are mildly higher for high SES students than they are for low SES students when controlling for student LEP and latent class. As a result, MLGMM $(\beta_{01highSES} - \beta_{01lowSES}) <$ MLLGM $(\beta_{01highSES} - \beta_{01lowSES})$ and MLGMM $(\beta_{11highSES} - \beta_{11lowSES}) <$ MLLGM $(\beta_{11highSES} - \beta_{11lowSES})$.

3.7 Characteristics of Cluster (Level-2) Data

My focus is on the school effect on the student's academic development. The level-2 cluster is schools and I add the level-2 covariate school SES (defined as previously stated) to remove the influence of students' selection to schools. As I mentioned , I identified selection to schools as one of the two sources of variation that may bias estimates of the school effect. Typically, researchers use the school SES as a categorical variable (e.g., High SES school vs. Low SES school). But, since I am proposing that school effect tends to be particularly more extreme for schools with very high percent levels or very low percent levels of students in FRL, this hypothesis can be better tested using the full range of FRL percent values, and the population of schools in my data has a range of between 15% and 99% of percent students receiving FRL, I propose to use the actual percent of students in FRL to take advantage of the variability in the data. As stated previously, I define schools with a high percent of students in FRL as very low SES schools and schools with a very low percent of students in FRL as very high SES. Following the criteria of the school district, I define schools with more than 50% of the students in FRL as Title I schools.

School SES was chosen as a covariate because schools receiving Title I funds are subject to NCLB sanctions while in most states other schools are not. These two group of schools (Title I schools and non Title I schools) differ with respect to multiple criteria such as the amount of resources available to them in the forms of school funding and teacher quality (Jackson, 2012). In addition, the Title I schools bear the added pressure of increased public scrutiny and the possibility of being labeled "failing schools" under NCLB due to the presence of students with challenging demographics. Schools in North Carolina are mainly evaluated based on the percent of students who are proficient in standardized tests in a given year; however, there is a strong association between students of low SES, ethnic minority status and poor school performance

(Peisner-Feinger, 2015). Since the current accountability metric in North Carolina does not account for growth or for relevant variables pertinent to growth, schools with high percent of low SES students are more likely to be categorized as failing schools based on the population they serve. Taking this information into consideration, I study the effect of the SES composition of the student body in schools as a strategy to capture schools' variability based on some student characteristics. The available evidence seems to indicate that the variability among schools based on some student characteristic (e.g., student SES or performance profile) effects school estimates (Yumoto, 2011). In addition to the misspecification of level-1 strata, having unaccounted-for variability between schools due to school SES can result in more extreme school estimates (level-2) for specific groups of schools with majority low and high SES students. More specifically, schools with a high proportion of low SES students may lead to school estimates (level-2) for low SES schools exaggeratedly higher in magnitude but negative in sign, and schools with a high proportion of high SES students may result with school estimates (level-2) for high SES schools exaggeratedly higher in magnitude and positive in sign (Yumoto, 2011).

3.7.1 School Variability

I use the school SES covariate to capture the cluster level variability due to school SES and its effects on school level estimates of student achievement (level-2) for schools with a given SES. The diversity of school SES mentioned previously allows for a large amount of variation at the cluster level. There is evidence indicating that when there is level-1 model misspecification (i.e., the LC was not specified when, in fact, it should be), variability at the cluster level (i.e., schools) leads to systematic bias in level-2 parameter estimates in multi-level models (Yumoto,2011). However, when MLGMM is used instead of simple MLLGM for the level-2 parameter estimates, reports from the literature indicate that the bias is greatly reduced in simulation studies (Yumoto, 2011; Muthén and Asparouhov 2009; Chen et al. 2010). Given this

interplay between model used and cluster (school) variability, I chose to incorporate students' SES composition within schools as a level-2 covariate. As described previously, the schools' SES composition represents the school SES covariate (a fixed variable). In a VAM context, if the level-2 data are conceptualized as representing the between-group level model (e.g., Equations 26-28), then school SES can be thought of as groups of schools causing variability based on their student SES composition.

Chen et al. (2010) only included two cluster levels (groups of schools based on high vs. low SES) for their level-2 covariate and half of the schools were high SES schools and half of the schools were low SES schools. However, in the evaluation of an effect of school in a VAM context, it is unrealistic to expect that all schools in the data will have equal proportions of schools in a given SES (high or low). Thus, unlike Chen et al. (2010), I include a range of schools each with its particular SES. The average reading EOG scores for the four years tend to be lower for students attending Title I schools and they tend to be particularly worse for students who attended Title I schools with a greater number of students in FRL. Schools with more than 87% of students in FRL have an average reading score for the four years of -0.50, schools with 50%-87% of students in FRL have an average reading score of 0.06, and schools with less than 50% of students in FRL have an average score of 0.493. These values are centered with an M of zero and SD unit of one.

3.7.2 Cluster Effects

The cluster effects, which describe the value added effect from a particular school, are the parameters of interest to me. The school estimates are the cluster effects and are represented by the level-2 parameters. When the level-2 model has no covariates, the parameters of interest are the following: γ_{0m0} and γ_{1m0}, where m represents the number of level-1 covariates (Equation 23). In the context of VAM analysis, the cluster-level effect represents the value-added effect,

γ_{1m0}, for a given level-1 covariate (e.g. student SES or LEP). When the level-2 model is conditioned on school SES, the parameters of interest are γ_{1m1} , where *m* represents the number of level-1 covariates. In the context of VAM analysis, the cluster-level effect represents the value-added effect of a school from a given school SES, γ_{1m1}, for a given level-1 covariate (e.g. student SES or LEP). Since my goal is to investigate the extent of changes in the level-1 on level-2 estimates, I apply several models from the most simple to the full expression of the proposed model. Initially, I apply four models to fit the data without covariates at level-2. First, I fit a traditional model without covariates at level-1 to the data. Second, I fit an additional traditional model to the data with student SES; next, I apply an additional traditional model with two level-1 manifest covariates, student SES and LEP, and finally I add the school level of analysis to the data. For the second case, I apply the same set of four models to the data but these models are different in that the level-2 has one covariate (school SES). This process is replicated when I apply an MLGMM model to the data.

I hypothesize that school level (cluster level) estimates derived from a conventional MLLGM are strongly more extreme than the estimates from a MLGMM whether level-2 has or does not have covariates. As a result, I assume that MLLGM |(γ_{010})| > MLGMM |(γ_{010})| and MLLGM |(γ_{110})| > MLGMM |(γ_{110})|for student SES; when the model incorporates LEP as level-1 covariate in addition to student SES, I also assume that MLLGM |(γ_{020})|> MLGMM |(γ_{020})| and MLLGM |(γ_{120})|> MLGMM |(γ_{120})|. In addition, when level-2 has school SES as a covariate, I hypothesize that school level estimates derived from a traditional MLLGM are more extreme than from a MLGMM such as MLLGM |(γ_{1m1})|> MLGMM |(γ_{1m1})|. I expect the cluster effect to vary based on the specific makeup of the cluster SES level. I hypothesize that schools with different proportions of students in a given SES causes variability; for instance, schools with a majority of students in low or high SES have caused variability in comparison with

schools with similar proportions of high and low SES students (Yumoto, 2011). Yumoto (2011) found that student heterogeneity (i.e., at level-2) contributes to the inflation of the cluster effect estimates, particularly when the model is miss-specified at level-1 (the latent variable at level-1 was not specified). For instance, schools with a high proportion of low performers would further lower negative cluster effect estimates and schools with a high proportion of high performers would increase higher positive effect estimates. As previously stated, in order to capture variability across schools, I included level-2 school SES effects as defined previously.

The school SES levels are specifically designed to evaluate the influence of variability of cluster in terms of the direction of biases (i.e., positive or negative), magnitude, and the precision of estimates; that is, since the purpose of a VAM is to estimate the impact of higher-level variables on the development or change in the first level variable (i.e., at the individual level), if there is no change, there can be no value-added effect estimated.

3.8 Model Fitting

My purpose is to model a traditional EVAAS model (MLLGM) and a MLGMM to assess the school effect on students growth. More specifically, I focus on the interpretation of the school effectiveness estimate after accounting for a latent class variable at level-1 and school variability at level-2 with a manifest covariate (school SES). I am particularly interested in assessing the interpretation of school effectiveness, given a school's SES. Following the approach of Chen et al. (2010), I will fit four models to infer the plausible number of LCs in the data: 1) traditional EVAAS model (i.e., MLLGM with LC unmodeled); 2) two alternative mixture models (i.e., MLGMM with two and three LCs); and, 3) an hypothesized mixture model (i.e., MLGMM with four LCs) to evaluate the effect of unmodeled LC (i.e. heterogeneity at student level-1) and the LC identification issues. I used *MPlus* 7 (Muthén & Muthén, 2016) to fit

these models. I use the growth profile in Table 4 as a criterion for the corresponding performance profiles in the mixture models.

3.8.1 Analysis of Results of Model Fitting

The identification of the presence of, and levels in, an LC variable in mixture models is based on more than one statistical index (Bauer & Curran, 2004; Nylund et al., 2007; Palardy & Vermunt, 2010). I utilized six indices – information criteria – to identify the model with the number of LC variables most representative of the data (Anderson, 2008). It is important for the model selection criteria to be robust because there may be multiple models that fit the data. The six information criteria that I used are:

- AIC (Akaike 1987)
- AIC3 (Bozdogan, 1993)
- AICc (McQuarrie & Tsai, 1998; after Akaike, 1987)
- BIC (Schwarz, 1978)
- BICB (Parlady & Vermunt, 2010)
- SABIC (Sclove, 1987)

All information criteria are defined as a function of the log-likelihood of the model; they differ in terms of the penalty each imposes based on the number of parameters estimated or sample size. Lower values of any information criterion indicate that the model for which it was computed fits the data better than do models with higher criterion values.

The following equations define each information criterion that I used:

$$AIC = -2logLL + 2P \tag{55}$$

$$AIC = -2logLL + 3P \tag{56}$$

$$AIC_c = -2LL + 2p\left[\frac{N}{N-p-1}\right] \tag{57}$$

$$BIC = -2LL + Plog(N) \tag{58}$$

$$BICB = -2LL + Plog(N_{classes}) \tag{59}$$

$$SABIC = -2LL + Plog\left(\frac{N+2}{24}\right) \tag{60}$$

where P is the number of estimated parameters, N is the sample size, and $N_{classes}$ is the number of classes. Prior work by scholars on research with GMMs indicates some details under which conditions specific information criteria may be more effective. Muthén and Asparouhov (2009) and Nylund et al. (2007) reported BIC to be one of the most effective information criteria to determine the correct number of latent classes with GMMs. Palardy and Vermunt (2010) found BICB to be more effective than BIC and AIC3 to be more effective than AIC. Anderson (2008) recommends against using BIC for multi-model selection exercises (see also Burnham & Anderson, 2002, and Yumoto, 2011), but its performance has been shown to be quite reliable and robust when used in simulations because the correct model is known to be among those in the model space (Anderson, 2008). However, Yumoto (2011) found that the AIC and AICc perform

slightly better for smaller cluster size samples (i.e., CS=20) while the BIC over-penalizes when small cluster sizes are present and the AIC3, SABIC and BICB perform similarly and well to identify MLGMM with the correct number of latent classes when cluster size was larger (i.e., CS=40).

I will specify seven models: one conventional unconditional level-1 growth model and six unconditional GMMs (also level-1). I impose several constraints on the mixture models. Models labeled A are GMMs with more restrictions (i.e., does not allow variation among the parameters within each class) while models labeled B are GMMs with less restriction (i.e., does allow some variation among some of the parameters within each class). The number before the letter of each model corresponds to the number of classes that are being specified, i.e., the conventional one class unconditional growth model is labeled 1, the growth mixture model with 2 classes is labeled 2, the growth mixture model with 3 classes is labeled 3 and the growth mixture model with 4 classes is labeled 4.

3.9 School Effects Analysis

I will process the parameter estimates I obtain from my *MPlus 7* analysis in SAS 9.3 (SAS Institute, 2009-2017). *MPlus* provides the estimates of individual growth parameters (level-1), student parameters (level-1), cluster level effects (i.e., intercepts and slope – level-2), and fit information (overall model). Group level effects for the MLGMM will be derived from the LC and the manifest variables and only from the manifest variables (LEP and student SES) for the MLLGM at level-1. I will then use a SAS program to convert group level effects (i.e., γ_{1m1}) to quintile rank for the estimates of both models, compute the variance of the group level effect, and construct 90% confidence interval (90% CI) over the estimates for that model. The computation necessary to assess the extent of the actual bias is not possible because we do not

know the true school level estimates, as shown in Equation 61. But I will discuss disagreement between methodologies (see next section).

$$B\left(\overline{\hat{\theta}}\right) = \overline{\hat{\theta}}_{est} - \theta_{true} \tag{61}$$

I will standardize the school value added scores based on a *t* statistic so that I may classify each school effect into a performance category with one of the most prevalent value-added models applied in practice, the EVAAS model used by the TAP (National Institute for Excellence in Teaching, 2009). To define what is sufficient evidence to conclude that a school requires special treatment, I test the null hypothesis that the school's performance is equal to the average performance in the school district (Solomon, White, Cohen, & Woo, 2007; Springer, Ballou, & Peng, 2008 and Schochet & Chiang, 2013).

In particular, I consider a performance measurement scheme that addresses the question, ‘‘Which schools performed particularly well or badly relative to the average school in the district?’’ Under this scheme, the considered null hypothesis is $H_0 : |\gamma_j - \bar{\gamma}_{\cdot}| = 0$, where $\bar{\gamma}_{\cdot}$ is the mean value of γ_j across all schools in the district. This testing approach will identify for special treatment schools for which the null hypothesis is rejected using a two-sided test, i.e., if a school’s value-added estimate is observed to be below or above the district average (Schochet & Chiang, 2013). I set a 5% risk threshold of committing Type I error with 40 degrees of freedom (N= 41-1, number of schools in the district minus 1) and use *t*= 2.021 as my cutoff (Schochet & Chiang, 2013). If the school value-added score is above 2.021, the school is considered to have grown their students above the district average and this school is labeled blue. On the other hand, if the school value-added score is below -2.021, the school is considered to have grown their students below the district average and this school is labeled red. Finally, if the school value-

added score is within the thresholds (2.021 and -2.021), the school is considered to have grown their students within the district average and this school is labeled green (TAP, 2009). Once the value-added scores are estimated for each methodological framework, I will use a repeated measures ANCOVA to test whether the value-added scores mean estimate for each method are statistically significantly different from each other. As a result, I will conduct two repeated ANCOVA tests: one to test whether the mean of the value-added scores across methodologies are different (when school SES is not specified at level-2) and the second to test whether the mean of the value-added scores across methodologies are different (when school SES is specified at level-2).

I will apply a similar procedure (i.e., repeated measure ANCOVA) to test whether the value-added scores' standard errors mean across methodologies are different. For this procedure I will convert the value-added scores' standard errors for each methodological framework into their ln form to maintain the normality assumption necessary to apply ANCOVA. Next, I will perform two repeated ANCOVA tests: one to test whether the mean of the converted value-added scores' standard errors across methodologies are different (when school SES is not specified at level-2) and the second to test whether the mean of the converted value-added scores' standard errors across methodologies are different (when school SES is specified at level-2).

3.9.1 Evaluation of Value-Added Scores Classification: Disagreement Rates

Every state has an accountability system to rank schools based on some criterion, such as how much schools grew their students, using value-added estimates to assess schools' performance. Type I error rate provides an upper bound on system error rates for individual schools. The Type I error rate (α) is the probability that based on *t* years of data, the hypothesis test will find that a truly average school performed significantly better or worse than average. Given α, the false positive error rate, FPR(*q*), is the probability that a school whose true

performance level is *q SD*s above or below average is falsely identified for special treatment (in either direction) (Schochet & Chiang, 2013). In this section, I address potential Type I error rates for measuring school performance due to the use of different methodological frameworks (MLGMM vs. MLLGM) when school SES is not and when school SES is specified at level-2. For the analysis, I classified the schools value-added scores into performance categories (i.e., blue, red and green), and derive results for different system error risk thresholds (1%, 5 % and 10%). Schochet and Chiang (2010) suggested that student heterogeneity is the key source of imprecision in estimating differences in value-added across schools thus if relevant sources of variation are not accounted for in the model, bias and imprecision are more likely to be introduced in school estimates. This in turn suggests that policymakers must carefully consider likely system error rates when using value-added estimates to make high-stakes decisions regarding schools. A smaller Type I error is considered better and it is expected that the better specified model (complex model or MLGMM) will yield a smaller number of schools in special treatment, thus when a more astringent risk threshold the disagreement rate between methodologies will be smaller, compared to the disagreement rate between methodologies when a less astringent risk threshold is used; for instance, the disagreement rate of a 1% risk threshold will be smaller than the disagreement rate of a 5% risk threshold and, in turn the disagreement rate when a 5% risk threshold is used will be smaller than the disagreement rate of a 10% risk threshold.

I will test the disagreement between the methods using a Kappa statistic because it accounts for the random agreement between methodologies (McHugh, 2012 and Yumoto, 2011). I will perform two tests: one to test the disagreement between the value-added scores classification derived by MLM and MLGMM when school SES is not specified at level-2 and the

second to assess the disagreement between the value-added scores classification derived by MLM and MLGMM when school SES is specified at level-2.

3.10 Summary of Methods

I am investigating the significance of various factors that may influence the biases in parameter estimates in the MLGMM context, – thereby integrating and refining the work of Muthén and Asparouhov (2009), Palardy and Vermunt (2010) and Yumoto (2011). In this chapter, I described the design features of my investigation of the effect of unmodeled heterogeneity at the individual level (level-1) on the precision and interpretation of estimation at higher levels in an MLGMM framework representing a generic VAM type analysis. Yumoto (2011) found that the estimates from MLGMM warrants further investigation in real data, particularly in the context of the teacher/school evaluation with VAM. One of my purposes in this work was to determine if improvement in estimation could be achieved in real data. My goal was also to investigate the impact on the school's (or cluster) effect estimates resulting from different proportions of poor students within a single school. I propose to use MLGMM as a tool to control bias and improve fairness in evaluation of school effect. Fairness in evaluation and policy making cannot be established if there is systematic bias in the estimates of any school's effect or effectiveness. I developed a variety of variables (i.e., emphasizing the accuracy of estimation of the school's effect or value-added after taking account the student's growth profiles, LEP, student SES, school SES and model type) in order to investigate the potential magnitude of bias and imprecision in schools' effect estimates resulting from ignoring the sources of variability (e.g. student characteristics omitted in the model and selection of students into schools) in the student population.

CHAPTER IV

MAIN STUDY RESULTS

In this chapter, I describe the results of my analysis as follows. In Section 4.1 I summarize the results of using a conventional one-level longitudinal single class MLM (without clusters) and I describe the use of the MLM to determine an appropriate growth curve function. In Section 4.2 I describe the results obtained when using a full single class two-level MLM model (with all available student covariates), including cluster and school level covariates. In Section 4.3 I summarize the results obtained by analyzing the data using a one-level mixture model as a way to illustrate key differences between single class and mixture models, in addition to determining the optimal number of classes in the model. In Section 4.4 I describe the results of using a full two-level mixture model with all available student and cluster covariates. In Section 4.5 I describe the results of the cluster effects when both frameworks are applied to the data as a way to illustrate key differences in the value-added scores classification of single-class and multiclass mixture models.

4.1 Model Identification 1: Longitudinal Multilevel Model Level-1

In this section I describe all the model specifications for a growth a model with regards to time scores, growth curve, residual variances and residual covariance for the latent variables or growth factors. I estimate two growth factors (initial status and growth rate), both of which are continuous latent variables and each of which has its own intercept and variance. Intercepts are means or averages for the initial status and growth rate. My assumption with respect to the variance is that all individual growth factors come from a common population and describe

differences between individuals. In addition, an intercept and a growth rate can be estimated for each student to account for variation across individuals. The mean of the intercept growth factor parameter represents the initial status and it is interpreted as the fixed part in the outcome variables at the time point where the time score (X_t=0, 1, 2 and 3 for four time points) is zero. The variance of the intercept growth factor is an estimate of the true variance (above and beyond the residual variance) of individuals at the time point with the time score of zero. In addition, I set the factor loadings for the initial status growth factor at 1 as part of the conventional parameterization of the outcome growth model in *Mplus*.

The mean of the growth rate factor parameter is interpreted as the average fixed part of the increase over individuals in the outcome variable for a time score increase of one unit. The variance of latent variable defining the slope is interpreted as the true variability of the growth rate across individuals. I set my time scores to 0, 1, 2 and 3 because my data contains four equidistant consecutive measures of reading scores and because I am assuming a linear growth function since I standardized the reading scores. Finally, the covariance between the latent variables describes their relationship. With regards to the other outcome parameters such as residual variances and residual covariances, residual variances represent time specific variation and measurement error and they can be considered to be unequal across time if the data determines this is the case. Residual covariances represent the relationships between time specific variation and measurement error sources of variation across time and independence is assumed.

I used ML growth model estimation under normality assumptions. With level-1 MLMs, model selection and modification were aided by fit indices such as the chi-square test (p>=0.05), RMSEA (<= 0.05), CFI/TLI (close to 1 or >= 0.95), and SRMR (<= 0.07). The LGM sources of misfit I tested included: the time scores for slope growth factor (to assess for the possibility of nonlinear growth), and the addition of relevant covariates to the model (FRL and LEP).

In order to make the best use of all available data and to avoid biases in parameter estimates, I assumed MAR with ML estimation for continuous outcome variables to model data with missing values (when individuals are not observed on all outcomes in the analysis). MAR implies that missing data points can be a function of the observed covariates and outcomes (e.g. the correlation between missing group and LEP(0.04) or the correlation between missing group and FRL (-0.37)). In my data, there was a moderate and negative correlation between high SES individuals and missing outcomes.

4.1.1 Basic Analysis

Preliminary descriptive sample statistical analysis of the data indicates the means of the outcome variables (standardized Reading scores for each year to a mean of zero and a SD of 1) were 0,-0.008, 0.016 and -0.017 for times 0, 1, 2, and 3 respectively. The variances were 1, 1.009, 1.086 and 1.018 for time 0, 1, 2, and 3 respectively. As expected, the correlations between the outcome variables were very high, ranging from 0.784 (between reading at time 1 and reading at time 3) to 0.997 (between reading at time 4 and reading at time 3)(Table 5).

Table 5. Correlation Matrix for Reading Scores for Times 1, 2, 3 and 4

Correlation Matrix

	Time 1	**Time 2**	**Time 3**	**Time 4**
Time 1	1			
Time 2	0.844	1		
Time 3	0.784	0.810	1	
Time 4	0.793	0.819	0.997	1

There were only some missing data points in the outcome variables but the percent of data present at each time point is better than the minimum required (10%) for good variance-covariance estimation convergence (100% of the data is present at time 0; 98% of the data is present at time 1; 97% is present at time 2 and 99% is present at time 3).

4.1.2 Unconditional Model (No Covariates Model)

Initially, I fit the LGM without covariates (also called the unconditional model) using fixed time scores (X_t =0, 1, 2, and 3 for the equidistant consecutive time measures). This step enabled me to determine the shape of the growth curve from the data. The estimated outcomes residual variances values are 0.197, 0.269, 0.056 and -0.077 for time scores 0, 1, 2 and 3 respectively (all were statistically significant). This first model indicated a negative residual variance when the time variable equaled 3 (also called a Heywood case), which suggested a nonlinear growth curve (Muthen, 2012). For this reason and the estimated Heywood case, I allowed time 3 to be estimated (to be free) to assess for a nonlinear growth function. This modification did not solve the negative variance problem. As a result, I imposed a more restricted model in which I set the residual variances of the outcome variables to be equal across time. Table 6 shows the resulting estimates with the following fit indices: chi-square of model fit of 9553. 806 with 8 d.f. (p=0.001), RMSEA = 0.596, CFI= 0.619, TLI=0.715, and SRMR=0.052. This alternative approach resolved the negative variance problem at time 3 thus I was able to assume a linear growth curve. However, this baseline model, in itself, did not fit the data very well as indicated by the chi-square of model fit (confirmed by the other model fit indices except SRMR), but this could be due to sensitivity of the test resulting from my large study population (N=3360) or to the lack of other variables (other than just the growth factors) needed to explain the outcome variables.

In the unconditional model, individual differences of the outcome variables are explained by the growth latent variables (also called growth factors): initial status and growth rate. Table 6 shows the estimates and standard errors for the longitudinal model latent variables (initial status and growth rate).

The model estimated value of the initial status mean (also called the mean of the initial status) is zero and the average growth rate is -0.005. The mean of the initial status and average growth rate are not statistically significant, suggesting that on average there was no sufficient evidence of change. The variance for the initial status was 0.897 (statistically significant) which suggests some variability in the initial status factor for individual differences. However, the variance for the growth rate is only 0.033; although statistically significant, this indicates a low variability in the growth rate factor for individual differences. The covariance between initial status and growth is negative (-0.032) and statistically significant, indicating that a high initial status is associated with a lower growth rate and a low initial status is associated with a higher growth rate.

Table 6. MLM One-Level Within Level Estimates

Level	Parameter	Estimate (Standard Error)		
		No Covariates	**FRL only**	**FRL & LEP**
Time	Residual Variance			
	Reading 1	0.097* (0.002)	0.097* (0.002)	0.097* (0.002)
	Reading 2	0.097* (0.002)	0.097* (0.002)	0.097* (0.002)
	Reading 3	0.097* (0.002)	0.097* (0.002)	0.097* (0.002)
	Reading 4	0.097* (0.002)	0.097* (0.002)	0.097* (0.002)
Student	Regression of Reading Intercept			
	on LEP			0.360* (0.037)
	on FRL		0.945* (0.030)	0.824* (0.032)
Student	Regression of Reading Growth Rate			
	on LEP			-0.012 (0.010)
	on FRL		-0.006 (0.008)	-0.002 (0.009)
Student	Mean/Intercept of Reading			
	for initial status	0.000 (0.017)	-0.394* (0.020)	-0.617* (0.030)
	for growth rate	-0.005 (0.004)	-0.003 (0.005)	0.005 (0.008)
Student	Variance/Residual Variance for Reading			
	Initial Status	0.897* (0.024)	0.680* (0.018)	0.660* (0.018)
	Growth Rate	0.033* (0.001)	0.033* (0.001)	0.033* (0.001)
Student	Covariance /Residual Covariance, reading initial status and growth rate	-0.032* (0.004)	-0.031* (0.004)	-0.030* (0.004)

*Statistically significant parameter estimate

Since I found significant differences between individuals with respect to the baseline performance as well as for growth, I used the next model to identify potential individual level predictors.

4.1.3 Add Covariates - FRL

I added covariates to growth models to help describe variation in reading scores at the initial time point and variation in the growth rate. FRL is a time invariant covariate which indirectly influences the outcome variables through the latent growth factors that I added to the model to explain the variability in the growth factors I found in the unconditional model (i.e., why some students start differently and why some students have different growth rates across time). I used this model to estimate the intercepts, residual variances, residual covariances for the growth factors and the effects of FRL on initial status and growth rate. I used the parameters of this model (intercepts and covariant coefficient) to estimate average growth factors and to estimate outcome means (which can be used for prediction purposes).

Table 2 shows the conditioned estimates and standard errors for the longitudinal model latent variables (initial status and growth rate) as well as the estimated effects of FRL (student level covariate) on the latent variables. The model estimated value of the initial status intercept is -0.394 and the growth rate intercept is -0.003; this is the average initial status and the average growth rate for students who receive FRL (respectively). The initial status intercept value is statistically significant but the growth rate intercept is not, suggesting that the influence of the FRL parameter on initial status is positive and significant (0.945); since I coded students who not received FRL as 1, this further suggests that students who do not receive FRL started higher (0.945) in the reading scores at time 0 than did the students who received FRL. However, the influence of the parameter FRL on growth rate was negative and non-significant (-0.006), suggesting that there was no significant difference in growth rate between students who received

or did not receive FRL and implies that those students without FRL remain at a higher achievement level on average. The variance for the initial status is 0.68 (statistically significant) and the variance for the growth rate is 0.033 (also statistically significant). The model estimated value for the outcomes residual variances is 0.097 (statistically significant). The residual covariance between the initial status and the growth rate is -0.031 (statistically significant), indicating that a high initial status is associated with a lower linear growth rate and vice versa.

Table 7 shows the corresponding model fit indices - chi-square of model fit (9580.102. d.f. = 10, p =0), CFI (0.632), TLI(0.632), RMSEA (0.534), and SRMR (0.044). These statistics suggest that this model does not fit the data well when FRL was added as a student level covariate.

4.1.4 Add Covariates FRL and LEP

I next added LEP (a time variant covariate) to the previous model but I treated as a time invariant covariant (I only considered the value at time 0) because I was interested in examining the differences at baseline between those students who were identified as LEP assumed that LCs will capture variability in the outcome variables across individuals.

Table 6 shows the conditioned estimated intercepts and standard errors for the longitudinal conventional model latent variables (initial status and growth rate) as well as the estimated effects of FRL and LEP (as student level covariates) on the latent variables. The model estimated value of the initial status intercept was negative and significant (-0.617) and the growth rate intercept was negative and non-significant (-0.005); these values represent the average initial status and the average growth rate of students who received FRL and who were LEP. The influence of the FRL parameter on initial status is still positive and significant (0.824) but somewhat diminished when compared that seen in the previous model (0.945). Since I coded students not receiving FRL as 1, this suggests that students who did not receive FRL started

higher (0.824) in the reading scores at time 0 than did the students who received FRL; the influence of the parameter FRL on growth rate remained negative and non-significant (-0.002), suggesting that there was still no difference in growth rate between students who received or did not receive FRL, implying that those students without FRL remained at a higher level of achievement on average. The influence of the parameter LEP on initial status was positive and significant; since I coded those not identified as LEPs as 1, this suggests then it can be interpreted that students who are not LEPs scored higher (0.36) in the reading scores at time 0 compared to students who identified as LEPs. Further, the influence of the parameter LEP on growth rate was negative (-0.012) and non-significant, suggesting that there was no difference in growth rate between students who were identified as LEP and those who were not so identified and implying those students not identified as LEP remain at a higher level of achievement on average.

Taken together, these statistics imply that not receiving FRL and not being LEP has a positive effect on students' reading scores initial status, but little if any effect on their growth rate. Further, the overall influence of FRL on the growth factors did not appear to change significantly with the addition of LEP to the model. The conditional variance for the initial status is 0.66 and significant, and the conditional variance for the growth rate is 0.033 and significant. The model estimated value for the outcomes residual variances is 0.097 and significant. The covariance between initial status and growth rate is negative and significant (-0.030). Table 7 shows the resulting model fit indices - chi-square of model fit (9585.889, d.f. = 12, p =0), CFI (0.633), TLI(0.572), RMSEA (0.487) and SRMR (0.038) – which indicate that this model still does not fit the data.

Table 7. MLM Fit Indices for One-Level

Fit Indices	**No Covariates**	**FRL only**	**FRL & LEP**
Number of free parameters	6	8	10
Chi-square model fit value	9553.806	9580.102	9585.889
degrees of freedom	8	10	12
p	0	0	0
RMSEA estimate	0.596	0.534	0.487
CFI	0.619	0.632	0.633
TLI	0.715	0.632	0.572
SRMR	0.052	0.044	0.038

In the next section, I describe my investigation of the need to account for observed variability of level-1 (Students) and level-2 (Schools) effects using the two-level MLLGM.

4.2 Model Identification 1: Two-Level Conventional MLM

In this section I consider the use of conventional growth modeling (an SEM analysis with two-level data) of individual-level and cluster-level data (i.e., repeated measures over grades for students nested within schools) to understand how schools vary in their ability to influence students' growth rate. The first level models the variation between students with respect to the

initial status and growth rate across time points; the second level models the variation across clusters or schools (or between variation). As a result, each growth factor (initial status and growth rate) is decomposed into uncorrelated within and between-cluster components, using subscripts *w* and *b* to represent within and between-cluster variation.

I analyzed this model using ML estimation in *Mplus* for unbalanced clusters and missing data. Since my data contained unbalanced clusters (clusters or schools with a different number of students) with missing data on the outcome variables, I chose to use full information ML estimation. In the first part of the model, I included all the student-level covariates (FRL and LEP), but the model lacks level-2 covariates (or unconditional level-2 only model). For the second modeling specification, I employed a key school-level covariate (school poverty index, measured as the percentage of the student body receiving full school lunch support).

4.2.1 Unconditional Two-Level MLM

My aim in my two-level growth modeling was the decomposition of the variability of students' initial status (i.e., intercept variance) into variability between students within a school (*iw* variation*)* and variability of average initial status between schools (*ib* variation). In addition, the analysis decomposes the variation of student growth into variation of growth between students within school (*sw* variation) and variability between schools with respect to average student growth (*sb* variation). As stated previously, here I modeled both the initial status (the intercept) and growth (the slope) as latent variables, *iw* and *sw*, which represented the level-1 variation in the intercept and slope growth factors across students, while the latent variables *ib* and *sb* represented the level-2 variation in the intercept and slope growth factors across schools. The decomposition of the latent variables explains how variation refers to covariates at student and school level. As a result, the decomposition into within and between components also occurs for the outcome variable residuals, consequently resulting into two sets of residual variances for

each level of analyses. The between residuals vary across schools and time while the within outcome variables' residuals vary across students and time.

The two-level growth model contains the same measurement specifications defined in the one level model with regards to time scores, growth curve, residual variances and residual covariance for the latent variables or growth factors. The same within and between time scores X_t are used on both levels. For this reason, I specified that the factor loadings for both between and within outcome variables' were the same (X_t = 0, 1, 2 and 3). The between portions of the outcome variable residual terms are equal to zero while the within part of the outcome variable residuals follows the initial restriction I imposed in the previous section. Further, I fixed the between outcome variables intercepts at zero because all mean values of the observed variables are predicted based on the values of the latent variables, thus the means of the growth factors are allowed to be free and estimated.

My population of schools comprises 41 schools (or clusters) with an average of 81.95 students per cluster; there are three schools with the lowest number of students (35, 36 and 37 students, respectively) and three schools with the largest number of students (one with 121 students and two with 127 students). The first part of my results here refers to the variation across students (or individual growth). Table 8 shows the within part (level-1) conditioned estimated effects on the latent variables of FRL and LEP with their respective standard errors for the unconditional (at level-2 only) multilevel longitudinal conventional model, as well as estimates of the latent variables residual variances and the latent variables residual covariance.

Table 8. MLM Two-Level Within Level Estimates

Level	Parameter	Estimate (Standard Error) No Covariates at Level-2	School SES
Time	Residual Variances		
	Reading 1	0.097* (0.002)	0.097* (0.002)
	Reading 2	0.097* (0.002)	0.097* (0.002)
	Reading 3	0.097* (0.002)	0.097* (0.002)
	Reading 4	0.097* (0.002)	0.097* (0.002)
Student	Regression of Reading Intercept		
	LEP	0.335* (0.038)	0.313* (0.038)
	FRL	0.581* (0.038)	0.545* (0.038)
Student	Regression of Reading Growth Rate		
	LEP	-0.016 (0.010)	-0.018 (0.011)
	FRL	-0.004 (0.010)	-0.007 (0.011)
Student	Residual Variances for Reading		
	Initial Status	0.606* (0.017)	0.606* (0.017)
	Growth Rate	0.032* (0.001)	0.032* (0.001)
Student	Residual Covariance, reading initial status and growth rate	-0.029* (0.003)	-0.029* (0.003)

*Statistically significant parameter estimate

The influence of the FRL parameter on initial status is positive and significant (0.581), but somewhat diminished when compared to the one-level model (0.824), likely because school level variation accounts for some of the initial status variability. Since I coded students not receiving FRL as 1, this suggests that students who did not receive FRL started higher (0.581) in the reading scores at time 0 than did the students who received FRL (conditioning on LEP status). In addition, the influence of the parameter FRL on growth rate remained negative and non-

significant (-0.004), which suggests that there was no difference in growth rate between students who did or did not receive FRL when controlling for LEP. The conditional influence of LEP status on initial reading score status was positive and significant (0.335). Since I coded students who were not identified as LEPs as 1, this suggests that students who are not LEPs started higher in the reading scores at time 0 compared to students who were identified as LEPs (conditioning on FRL).

The conditional influence of the parameter LEP on growth rate remained negative (-0.016) and non-significant which suggests that there was no difference in growth rate between students who were, or were not, identified as LEPS (conditioning on FRL). Taken together, the average initial status for students who do not receive FRL are higher than the initial status for students who receive FRL, and there were no differences in growth rate, which implies that those students who did not receive FRL remained at a higher level of achievement on average. The same conditions appear to hold true for LEP status – non-LEP students began at a higher reading score status and remained at a higher level of achievement. However, adding clusters (school level) covariates to the model did not change significantly the student-level covariates with respect to growth rate status from those in the one-level longitudinal model, but the covariance analysis suggests something different. The residual variance for the initial status is 0.606 and significant, and the residual variance for the growth rate is 0.032 and significant. The model estimated value for the outcomes residual variances is 0.097. The covariance between initial status and growth is negative (-0.032) and significant which indicates (contrary to my previous findings) that a high initial status is associated with a lower growth rate and a low initial status is associated with a higher growth rate and significant. The covariance between initial status and growth rate is negative and significant (-0.029).

In the level-2 portion of the analysis, I describe the variation of the between level (school) initial status and growth rate across schools (or school growth). Table 9 shows the latent variables average estimates and standard errors for the unconditional (level-2 only) two-level longitudinal conventional model, latent variables variances and latent variables covariance. The model estimated mean initial status for schools with zero percent of students who receive FRL is negative and significant (-0.527) and the average school growth rate intercept is positive and non-significant (0.010). The variance for the school initial status is 0.081 and significant, and the variance for the school growth rate is 0.001 and significant. The between covariance of school initial status and school growth rate differs from the within pattern covariance (small negative association). The residual covariance of the between school initial status and school growth rate is negative and non-significant (-0.001) and thus implies that there is no association between initial status and growth for schools.

Table 10 shows the model fit indices for the level-2 unconditional MLM - chi-square of model fit (9614.038, d.f. = 19, p=0), CFI (0.592), TLI(0.571), RMSEA (0.388) and SRMR (within=0.039, between=0.011) – all of which suggest that this model fits the data poorly.

Table 9. MLM Two-Level Between Level Estimates

Level	Parameter	Estimate (Standard Error) No Covariates at Level-2	Estimate (Standard Error) School SES
School	Regression of Reading Intercept on School SES		-0.839* (0.103)
School	Regression of Reading Growth Rate on School SES		-0.017 (0.025)
School	Means/Intercept of Reading for Initial Status	-0.527* (0.055)	0.057 (0.085)
	for Growth Rate	0.010 (0.010)	0.023 (0.021)
School	Variances/Residual Variances for Reading for Initial Status	0.081* (0.021)	0.024* (0.007)
	for Growth Rate	0.001* (0.00)	0.001* (0.00)
School	Covariance/Residual Covariance, reading initial status and growth rate	-0.001 (0.002)	-0.002 (0.001)

*Statistically significant parameter estimate

Table 10. MLM Fit Indices for Two-Level

Fit Indices	**No Covariates at Level-2**	**School SES**
Number of free parameters	13	15
χ^2 square model fit value	9614.038	9630.318
degrees of freedom	19	21
p	0	0
RMSEA	0.388	0.369
CFI	0.592	0.593
TLI	0.571	0.535
SRMR		
value within	0.039	0.039
value between	0.011	0.009

4.2.2 Conditional Two-Level MLM (School SES)

In this section, I focus on understanding between-level variation. More specifically, here I assess how schools vary with respect to their ability to influence students' growth rate when I include in the model a school-level covariate "School SES". All other aspects of the between-level model were specified in the same way as the previous model. I set school-level outcome

variables' residual variances (equal to zero), between outcome variables' factor loadings (X_t = 0,1, 2 and 3), and between outcome variables' intercepts (equal to zero). First, I will discuss the results concerning the variation across students. Table 8 shows the within part (level-1) conditioned estimated effects of FRL and LEP on the latent variables for the conditional (at level-2) conventional MLM, latent variables residual variances and latent variables residual covariance. The influence of the FRL parameter on initial status is positive and significant (0.545). However, this value is somewhat smaller when compared to the unconditional two-level model (0.581), likely because school SES accounts for some of the initial status variability. Since I coded students not receiving FRL as 1, this suggests that students who did not receive FRL started with higher reading scores at time 0 than the students who received FRL (when conditioning on LEP).

The influence of the parameter FRL on growth rate remained negative and non-significant (-0.007), suggesting that there was no difference in growth rate between students who received FRL or did not receive FRL (when conditioning on LEP). Similar to FRL, the influence of the parameter LEP on initial status was positive and significant (0.313), when conditioning on FRL. Since I coded students not identified as LEPs as 1, this suggests that students who were not LEPs scored higher in reading at time 0 compared to students who were identified as LEPs. The influence of the parameter LEP on growth rate remained negative (-0.018) and non-significant (when conditioning on FRL), which suggests that there was no difference in growth rate between students who were, or were not identified as LEPS.

Taken together, this implies that not receiving FRL and not being LEP has a positive effect on reading scores initial status, but these two covariates do not seem to have any effect on students' linear growth rate, which further suggests that non-FRL students and non-LEP students remained at a higher level of achievement than their FRL-receiving or LEP counterparts. The observant reader will notice that the interpretation of student-level covariates when school SES is

included in the model has not changed from the previous two-level unconditional model or the one-level model. The residual variance for the initial status is 0.606 and significant, and the residual variance for the growth rate is 0.032 and significant. The residual covariance between initial status and growth rate is negative and significant (-0.029).

In the level-2 portion of the analysis, I describe the variation of the between level (school) initial status and growth rate across schools (or school growth). Table 9 shows the latent variables intercept estimates and standard errors for the conditional (level-2) two-level longitudinal conventional model, the effect of level-2 covariate "school SES" on latent variables, latent variables residual variances and latent variables residual covariance. The model estimated school mean initial status (intercept) when school SES equals zero is positive and non significant (0.057) and the average school growth rate (intercept) when school SES equals zero is positive and non-significant (0.023). The influence of the school SES parameter on school initial status is negative and significant (-0.839).

Since I specified school SES to be the inverse of the percentage of students who received FRL, a unit increase in percent of students in FRL results in a decrease of the school average reading score initial status of 0.839. In addition, the influence of the variable school SES on school growth rate was negative and non-significant (-0.017), which suggests that there was no difference among schools with different SES with respect to their average growth rate. In summary, these results imply that having a higher proportion of low SES students have a negative effect on school average initial status on reading scores, but does not seem to have any effect on schools' average growth rate. The residual variance for the school initial status is (0.024) and significant, although somewhat reduced from the previous model (0.081), and the school residual variance for the growth rate is (0.001) and significant. The residual covariance between school initial status and school growth rate is negative and non-significant (-0.002) which indicates that

there is no relationship between school average initial status and school average growth rate. Table 10 shows the model fit indices -chi-square of model fit (9630.318, d.f. = 21, p=0), CFI (0.593), TLI(0.535), and RMSEA (0.369) and SRMR (within=0.039, between=0.009) – which suggest that this model does not fit the data.

4.3 Model Specification 2: Growth Mixture Model Analysis Results Level-1

In this section, I apply a one-level regression growth mixture model to the reading scores. My aim for the mixture modeling approach presented in this section is to address the unobserved variability of level-1 effects. In the previous model specification, I assumed that the variation between students with respect to the growth factors was unobserved (r_{0ij} and r_{1ij}). However, using a mixture modeling approach I assume that part of the variation of the r_{0ij} and r_{1ij} can be explained by the existence of underlying LCs. More specifically, each LC represent a given trajectory class (a subpopulation in the data). In other words, GMMs allow heterogeneity with respect to growth functions in which different classes correspond to different growth shapes (each population has its own initial status and growth rate).

I imposed several constraints on the model. These restrictions were necessary to identify the more advanced models (MLGMM) in later sections. I was limited by the number of parameters that can be freely estimated due to the smaller number of clusters (i.e., 41 schools). For this reason, I attempt to fit six mixture models: model 1A, model 1B, model 2A, model 2B, model 3A and model 3B. In the "A" sequence I do not allow variation within each class but in the "B" sequence I do allow some variation within each class (see Table 11). I parameterize these models using the same specifications described previously and I continue to apply the restriction of equal residual variances across time as it was done for the conventional regression MLM.

Table 11. Restrictions per Model Specification

		Variation per Latent Class				
Model	Classes Number	Residual Variances time1-4	Growth Factors	Residual Variances for Growth Factors	Residual Covariance between Growth Factors	Level-1 parameters on growth factors effects
1	1	No	No	No	No	No
1A	2	No	Yes	No	No	No
1B	2	Yes	Yes	No	No	No
2A	3	No	Yes	No	No	No
2B	3	Yes	Yes	No	No	No
3A	4	No	Yes	No	No	No
3B	4	Yes	Yes	No	No	No

4.3.1 Growth Mixture Modeling with Latent Classes: Models A and B

As in the conventional model 1, Models A and B explain the reading outcomes through the latent variables (initial status and growth rate), but additionally models A and B contain specifications for trajectory LCs at level-1. In this way, the model captures heterogeneity by using both categorical (LCs *c*) and continuous latent variables (growth factors). Thus the LCs describe different class-specific average growth curves (i.e., a classes average initial status and a classes average growth rate), and within each class, individual students' growth curve differences

from their respective class average growth curve. I assess each model initially in its unconditional form to determine the number of meaningful classes. Model 1 is the unconditional conventional model; Models 2, 3, and 4 extend that base model by adding (respectively) two, three, or four qualitatively different growth curves (i.e., LCs).

4.3.2 Multinomial Logistic Regression Parameterization

Here I model the probability that an individual falls in a given class as a function of the given covariates (if added) and the growth factors using a multinomial logistic regression model. Some variables may be more related to class than others and the classification of LC membership may potentially vary when model specification changes (e.g. adding covariates). By using a continuous factor to represent variation and co-variation, I need three random intercepts because the probabilities must add to 1, and, with four hypothesized LCs (K=4), I can express these multinomial random intercepts. When I add student-level covariates to the model, I allow the covariates to have a direct influence on each of the intercepts and, by implication, all probabilities.

4.3.3 Reading Scores Unconditional Model Comparisons

To choose the optimal number of classes, I use the unconditional level-1 growth mixture model. Comparisons of model fit that have different number of classes is typically accomplished by using the BIC (Yumoto, 2011), but I also assess the models using other fit indices (BICB, SABIC, AIC, AIC3 and AICc); the lower value of a fit index, the better the model, thus I increase the number of classes until I find the fit indices minimum. Table 12 shows the number of parameters, number of classes and fit indices for the models I considered.

Table 12. GMM Model Comparisons Class Fit

Model	N classes	Log	N pars	BIC	SABIC	BICB	AIC	AIC3	AICc
1	1	-11022	6	22093	22074	22044	22056	22062	22056
1A	2	-10828	9	21728	21700	21658	21673	21682	21673
1B	2	-10660	10	21402	21370	21324	21341	21351	21341
2A	3	-10755	12	21607	21568	21515	21533	21545	21533
2B	3	-10558	14	21230	21186	21123	21145	21159	21145
3A	4	-10715	15	21551	21503	21438	21459	21474	21459
3B	4	-10482	18	21109	21052	20974	20999	21017	20999

There are two notable trends in Table 12. First, the "B" sequence models consistently show a better fit than the corresponding "A" sequence models, which strongly suggests that the more general models (less restricted models) have the best fit indices-values. Second, in general, within each sequence, the model fit improves concomitantly with the number of classes, thus the model 3B fits better than 2B, which in turn fits better than model 1B (and the same holds true for the "A" sequence). Thus of all the models, model 3B (4-class, less restricted) has the best (lowest) fit indices-value, suggesting that four LCs are optimal.

4.3.4 Conditional Models Comparisons

I used the usual log-likelihood ratio to compare the model fit of two models that have the same number of classes and are nested. Model 1 in Table 13 is the unconditional four-class mixture model for the one-level model, Model 2 includes the student-level conditioned with the covariate FRL, and Model 3 includes the student-level conditioned with two covariates FRL and LEP. Among the 4-class models, Model 3 has the best (lowest) Log-likelihood value, but the improvement in fit between Model 2 and Model 3 is less than that for Models 1 and 2 (suggesting diminishing returns as the number of classes increases). These results confirm the value of my

choice to define FRL as a covariate at level-1 as part of the model specification. Comparing the three versions of the four-class models, I choose model 3 on the basis of Log-likelihood.

Table 13. MLGMM Level-1 Model Comparisons Fit

Model	Number of classes	Log-likelihood	Number of parameters	BIC	SABIC	AIC
1	4	-10482	18	21109	21052	20999
2	4	-10013	23	20214	20141	20073
3	4	-9962	28	20152	20063	19981

I also found that adding student level covariates (FRL and LEP) did not affect the LCs classification, thus the population represented by each class did not differ much when I added covariates to the model. The unconditional four-class model (Model 1, Table 14) classified 38% of the students as HP, 34% as PLP, 25% as LP, and 3% as S; the four-class model with only FRL as the student-level covariate (Model 2, Table 15) classified 38% of the students as HP, 32% as PLP, 28% as LP, and 2% as S; the four-class model (Model 3, Table 16) with all level-1 covariates classified the students identically to that of Model 2. Unlike the conventional models, adding the student-level covariates (FRL and LEP) greatly improved the model fit (especially adding FRL).

Table 14. MLGMM Class Membership Unconditional

Class	Initial Status	Growth Rate	Probability
Class 1 (LP)	0.18	-0.03	0.25
Class 2 (S)	-0.75	0.59	0.03
Class 3 (HP)	0.72	0.00	0.38
Class 4 (PLP)	-0.87	-0.04	0.34

Table 15. MLGMM Class Membership FRL only

Class	Initial Status	Growth Rate	Probability
Class 1 (LP)	-0.07	-0.03	0.28
Class 2 (S)	-0.85	0.64	0.02
Class 3 (HP)	0.35	-0.01	0.38
Class 4 (PLP)	-1.05	-0.03	0.32

Table 16. MLGMM Class Membership FRL and LEP

Class	Initial Status	Growth Rate	Probability
Class 1 (LP)	-0.24	-0.02	0.28
Class 2 (S)	-1.04	0.67	0.02
Class 3 (HP)	0.18	0.01	0.38
Class 4 (PLP)	-1.20	-0.02	0.32

4.3.5 Growth Mixture Full (FRL and LEP) Model Results

In this section, I discuss the estimates for the mixture four-class model (Model 3, Table 17). I draw the conclusions from the four-LC model with all level-1 covariates that are somewhat different from those I drew from the conventional one-class model with all level-1 covariates. These four LCs are always ordered from high to low for the reading achievement initial status: 0.177 (class HP, 38%), -0.24 (class LP, 28%), -1.039 (class S, 2%) and -1.199 (class PLP, 32%) (all of which were statistically significant); note that these values are in SD units, where an SD unit is a mean difference of 0.18. The growth rate for these four classes are: 0.01 for class HP, -0.02 for class LP, 0.67 for class S and -0.02 for class PLP; only the growth rate for class S is statistically significantly different from 0 (unlike the conventional model, where the growth rate intercept was not). I restricted the residual variances for each of the LCs' initial status and growth rate to be equal (0.33 (S.E. 0.027) and 0.019 (S.E. 0.001) respectively).

Table 17. GMM Within Estimates FRL and LEP

Level	Parameter	Estimate (Standard Error)			
		Class 1	Class 2	Class 3	Class 4
Time	Residual Variance				
	Reading 1	0.027* (0.004)	0.595* (0.084)	0.103* (0.008)	0.137* (0.009)
	Reading 2	0.027* (0.004)	0.595* (0.084)	0.103* (0.008)	0.137* (0.009)
	Reading 3	0.027* (0.004)	0.595* (0.084)	0.103* (0.008)	0.137* (0.009)
	Reading 4	0.027* (0.004)	0.595* (0.084)	0.103* (0.008)	0.137* (0.009)
Student	Regression of Reading Intercept				
	on LEP	0.254* (0.071)	0.254* (0.071)	0.254* (0.071)	0.254* (0.071)
	on FRL	0.515* (0.051)	0.515* (0.051)	0.515* (0.051)	0.515* (0.051)
Student	Regression of Reading Growth				
	on LEP	-0.027* (0.009)	-0.027* (0.009)	-0.027* (0.009)	-0.027* (0.009)
	on FRL	0.014 (0.010)	0.014 (0.010)	0.014 (0.010)	0.014 (0.010)
Student	Intercept of Reading				
	for Initial Status	-0.240* (0.093)	-1.039* (0.153)	0.177* (0.027)	-1.199* (0.065)
	for Growth Rate	-0.018 (0.010)	0.666* (0.096)	0.013 (0.001)	-0.019 (0.017)
Student	Residual Variance for Reading				
	Initial Status	0.330* (0.027)	0.330* (0.027)	0.330* (0.027)	0.330* (0.027)
	Growth Rate	0.019* (0.001)	0.019* (0.001)	0.019* (0.001)	0.019* (0.001)
Student	Residual Covariance, reading initial status and growth rate	-0.022* (0.007)	-0.022* (0.007)	-0.022* (0.007)	-0.022* (0.007)

*Statistically significant parameter estimate

In Table 17, Class 3 appears to represent HP students (highest start with no growth), Class 1 represents as LPs (low start with no growth), and Class 4 represents the PLPs –(the lowest start with no growth). However, Class 2 (which should represent the S group) did not quite fit the initial hypothesized parameters; I had expected that LPs and Ss would have similar starting points, with S progressing at a strong positive growth rate but LPs would have zero growth. However, I found that the S group's initial status is similar to that of PLPs' with a very strong positive growth rate (as hypothesized); despite this discrepancy from my expectations and hypothesis, I continue to refer to class 2 as the S class. Each trajectory class has its own residual variances across time: 0.027 for class 1, 0.595 for class 2, 0.103 for class 3, and 0.137 for class 4.

At the student-level, I use linear regression to relate the growth factors to the covariates FRL and LEP. The regression of the initial status and growth rate on FRL and LEP do not vary across classes. This restriction was necessary for me to identify a more complex model (MLGMM) in later sections. I evaluated the initial status for these four classes for students who do not receive FRL and who are not LEPs. Both covariates' influence on initial status intercept were positive and statistically significant (0.254 for LEP and 0.515 for FRL). Since I coded students who did not receive FRL as 1, this suggests that students who did not receive FRL had a higher intercept in the reading scores at time 0 than did the students who received FRL. In addition, the influence of the variable FRL on growth rate was negative and non-significant (0.014), which suggests that there was no difference between students who received or did not receive FRL in their growth rate. Since I coded non-LEP students as 1, the result for LEP suggests that students who were not LEP had a higher initial status (0.254) in the reading scores at time 0 than did the students who were LEP; for the same reason, the negative but statistically significant (-0.027) influence of LEP status on growth rate suggests that students who are not LEP had a more negative growth rate.

Taken together, these results indicate a disadvantage for students who receive FRL and who are identified as LEPs at the start of the growth curve. However, students' growth rate was only related to LEP (and not to FRL). So far, with respect to the influence of the student-level covariates on initial status interpretation, these results are similar to those I obtained from the conventional one-single class model. However, the impact of LEP on growth rate in this model is different from that which I observed in the conventional model, where no LC distinction was made. In this way, the multilevel growth mixture modeling results imply that FRL does not influence the reading growth rate but that LEP does although indirectly through its influence on achievement trajectory class, which in turn influences student growth.

To understand who resides in these classes, what type of individuals they are and what covariates are related to the probability of belonging to a given class, I describe the results of the multinomial logistic regression of LC membership on the covariates. I found that level-1 covariates have a significant influence on c in the sense that a high value resulted in a higher probability of being a member of the class. I found that membership in class 1 (LP) and class 3 (HP) was predicted by not receiving FRL (0.860 for class 1 and 1.266 for class 3) while that of class 2 (S) was predicted by not being LEP (0.931). In other words, when PLP is the reference group, FRL had a significant influence on the probability of being a member of the class with a good reading achievement (class 3 HP) trajectory and also of being a member of the class with a poor reading achievement (class 1-LP) trajectory in Grades 3 through 6, while LEP had a significant influence on the probability of being a member of the class with strong growth reading achievement (class 2-S) trajectory in Grades 3 through 6 (see Table 18).

Table 18. GMM Multinomial Estimates FRL and LEP

Parameter	Estimate	S.E	EST./S.E.
C1 ON			
LEP	0.107	0.201	0.533
FRL	0.86*	0.19	4.519
C2 ON			
LEP	0.931*	0.417	2.233
FRL	-0.452	0.536	-0.843
C3 ON			
LEP	0.545	0.337	1.614
FRL	1.266*	0.218	5.81

*Statistically significant parameter estimate

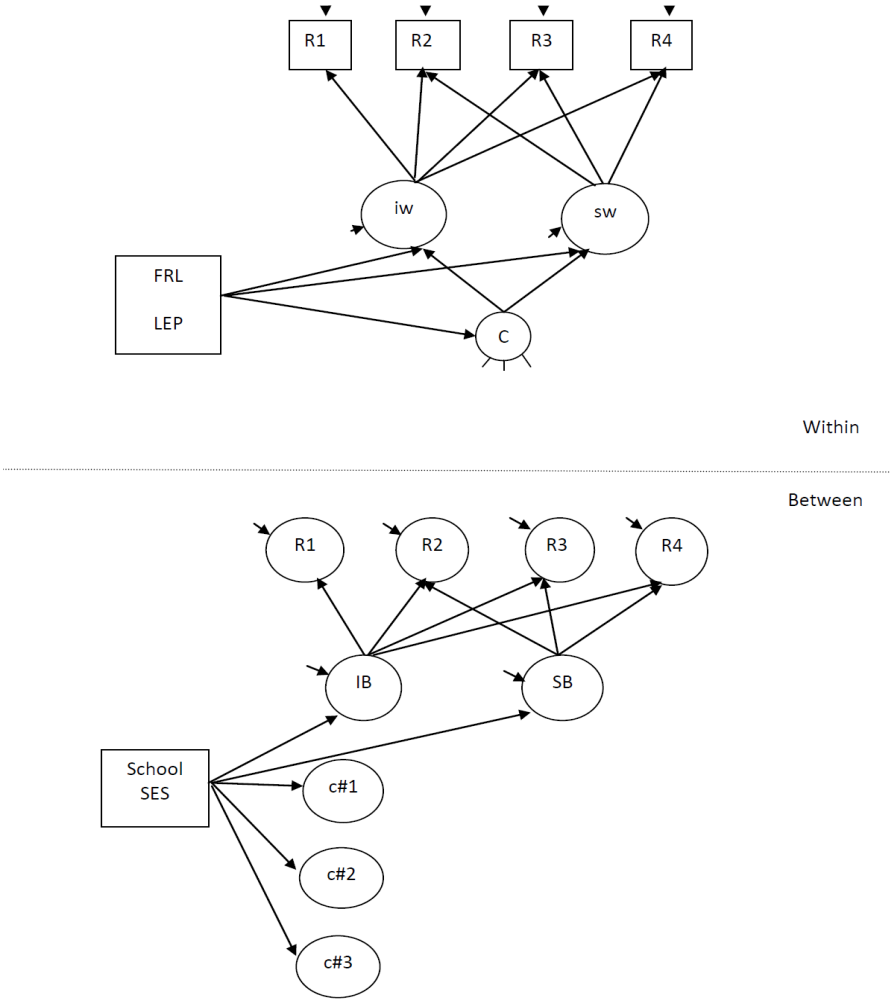
4.4 Model Specification 2: MLGMM Analysis Results Level-2

The growth model of Section 4.2 contains an assumption that all individuals come from one and the same population. This is seen in equations 31 and 32 where there is only one set of parameters for β. However, similar to the one-level regression mixture results of Section 2.3,

there may be unobserved heterogeneity in the data corresponding to different subpopulations of reading developmental trajectories. This type of heterogeneity is captured by LCs, i.e. finite mixture modeling. In addition to the regular GMM , I specified another level (level-2) of analysis for this model, i.e., I added the school-level (or cluster) to this model to estimate the contribution of schools to student learning. Figure 1 shows the model diagram for the two-level MLGMM for the reading data with all the between and within covariates.

In the within (student-level) part of the model, the LC variable *c* influences the growth factors *iw* and *sw*. The lack of broken arrows from *c* to the arrows from the set of covariates to the growth factors indicates that the covariate is constrained to be equal across the LCs. The three short lines for *c* indicate random intercepts (similarly to the conventional on-class MLM). These random intercepts are continuous latent variables that modeled using the between (school-level) portion of the model. The between-level circles for the dependent variables reading1-reading4 (R1-R4) have intercepts. I define the between school portion of the growth factors, the initial status and the growth rate as random effects. In other words, the growth factors vary across schools in addition to the variation across students within schools. Since I estimated four LCs , there are three random intercepts for *c*, labeled c#1, c#2 and c#3 (class 4 is the reference group). In the between-level model, as mentioned previously, I specified the percent of students in FRL as a measure of the school poverty index. In the final model (MLGMM) I specify the combined multilevel influence of the between and within level covariates to the classification of the trajectory classes. With this model specification, I define the between-level covariate relationship with the LC intercepts such that it can influence the random intercept value of any given class, which in turn, it makes it more likely to belong to any given trajectory. All mixture models are estimated using ML because it uses all available information and it is better for estimation of higher order moments (such as skewness and kurtosis) (Muthen, 2012).

Figure 5. Four-Class MLGMM with LEP, FRL and School SES



4.4.1 Unconditional Two-Level MLGMM

To understand which students are in a given development class, I describe the relationship between the within and between covariates, and the LCs. Table 19 shows the estimates for the multinomial logistic regression of *c* on the student level covariates which provides a sorting of the observed trajectories into four LCs: HP, LP, S and PLP, in which the reference class is PLP (class 3).

Table 19. MLGMM Multinomial Estimates without School SES

Parameter	Estimate	S.E.	Est./S.E.
Within level			
S On			
LEP	1.055*	0.498	2.118
FRL	-0.817	0.594	-1.376
HP On			
LEP	0.543	0.423	1.283
FRL	1.016*	0.243	4.181
LP On			
LEP	0.129	0.251	0.514
FRL	0.811*	0.188	4.312

The results indicate that the probability of membership in class 1 (S), relative to the poorly developing reference class 3 (PLP) are statistically significantly (1.055) if the student was not LEP (FRL had no influence). The probability of membership in class 2 (HP), again relative to PLP, is statistically significantly (1.016) for non-FRL individuals (LEP has no influence). The probability of class 4 membership (LP), relative to the poorly developing reference class 3 (PLP) is statistically significant (0.811) for non-FRL individuals not receiving FRL (compared to FRL

students) , but LEP had no influence. Table 20 pertains to the within-level estimates and shows results for the influence of the student-level covariates on the growth factors intercepts for each class, residual variance for the growth factors for each class and residual covariance for the growth factors for each class.

Both of the covariates' influences on the initial status mean are positive and statistically significant (0.229 for LEP and 0.361 for FRL), which suggests (since I coded non-FRL students as 1) that non-FRL students had a higher initial status average in the reading scores at time 0 than their FRL counterparts. However, the influence of the FRL on growth rate was negative and not significant (0.012), which suggests that there was no difference with respect to growth rates between FRL and non-FRL students. Similarly, since I coded non-LEP students as 1, those students had a higher initial status mean (0.229) in the reading scores at time 0 than did the LEP students.

The influence of LEP status on growth rate is negative and statistically significant (-0.029), which suggests (since I coded non-LEP students as 1) that the non-LEP students had a more negative growth rate compared to LEP students. In addition, the residual variance for the student initial status is 0.3 and significant, and the residual variance for the student growth rate is 0.018 and significant. The residual covariance between the growth factors is (-0.018) and significant, which indicates that students with higher initial status had lower growth rates or that students with lower initial status had higher growth rates. This result is similar to that of the one-class two-level MLM without school SES as level-2 covariate from section 4.2.1.

Table 20. MLGMM Within Estimates without School SES

Level	Parameter	Estimate (Standard Error) Class 3	Class 2	Class 1	Class 4
	Within level				
Time	Residual Variances				
	Reading 1	0.140* (0.010)	0.108* (0.008)	0.619* (0.066)	0.029* (0.004)
	Reading 2	0.140* (0.010)	0.108* (0.008)	0.619* (0.066)	0.029* (0.004)
	Reading 3	0.140* (0.010)	0.108* (0.008)	0.619* (0.066)	0.029* (0.004)
	Reading 4	0.140* (0.010)	0.108* (0.008)	0.619* (0.066)	0.029* (0.004)
Student	Regression of Reading Intercept				
	LEP	0.229* (0.111)	0.229* (0.111)	0.229* (0.111)	0.229* (0.111)
	FRL	0.361* (0.073)	0.361* (0.073)	0.361* (0.073)	0.361* (0.073)
Student	Regression of Reading Growth Rate				
	LEP	-0.029* (0.011)	-0.029* (0.011)	-0.029* (0.011)	-0.029* (0.011)
	FRL	0.012 (0.012)	0.012 (0.012)	0.012 (0.012)	0.012 (0.012)
Student	Residual Variances for Reading				
	Initial Status	0.300* (0.026)	0.300* (0.026)	0.300* (0.026)	0.300* (0.026)
	Growth Rate	0.018* (0.001)	0.018* (0.001)	0.018* (0.001)	0.018* (0.001)
Student	Residual Covariance, reading initial status and growth rate	-0.018* (0.007)	-0.018* (0.007)	-0.018* (0.007)	-0.018* (0.007)

Table 21 pertains to the between-level estimates and gives results for the average initial status *ib* and average growth rate *sb* for each LC, the variance between the growth factors for each LC and the covariance between the growth factors for each LC.

Table 21. MLGMM Between Estimates without School SES

Level	Parameter	Estimate (Standard Error) Class 3	Class 2	Class 1	Class 4
	Between level				
School	Means of Reading				
	for Initial Status	-1.156* (0.106)	0.186 (0.121)	-0.879* (0.215)	-0.177 (0.100)
	for Growth Rate	-0.001 (0.025)	0.005 (0.027)	0.701* (0.095)	-0.013 (0.011)
School	Variances for Reading				
	for Initial Status	0.069* (0.014)	0.069* (0.014)	0.069* (0.014)	0.069* (0.014)
	for Growth Rate	0.001* (0)	0.001* (0)	0.001* (0)	0.001* (0)
School	Covariance, reading initial status and growth rate	0 (0.002)	0 (0.002)	0 (0.002)	0 (0.002)

The four LCs are ordered from high to low reading achievement average initial status: 0.186 (class 2, 37%), -0.177 (class 4, 31%), -0.879 (class 1, 2%) and -1.156 (class 3, 30%). Here, the average initial status for classes 3 and 1 are statistically significantly different from zero. The average growth rate for these four classes are: 0.005 for class 2, -0.013 for class 4, 0.701 for class 1 and -0.001 for class 3. Here, only the average growth rate for class 1 is statistically significantly different from 0. The results for the classification of the latent trajectories are similar to the results of the previous section (one-level GMM model).

Table 22 shows the results for each LC probability classification. The variance for the school initial status is 0.069 and statistically significant, and the variance for the school growth rate is 0.001 and significant. The covariance between the growth factors show that the two growth factors are not correlated so that when school SES equals zero the reading initial performance mean in a school is not associated with the school growth rate average. This result is similar to the one-class two-level MLM without school SES as level-2 covariate from section 4.2.1.

Table 22. MLGMM Class Membership without School SES

Class	Initial Status	Growth Rate	Probability
Class 2 (HP)	0.186	0.005	0.37
Class 4 (LP)	-0.177	-0.013	0.31
Class 1 (S)	-0.879*	0.701 *	0.02
Class 3 (PLP)	-1.156*	-0.001	0.30

4.4.2 Conditional Two-Level MLGMM

Table 23 shows the estimates for the multinomial logistic regression of *c* on the within and between covariates, which provides a sorting of the observed trajectories into four LCs (PLPs are the reference class). Figure 6 shows the estimated trajectory classes for reading achievement in Grades 3-6 for this model and Figure 7 shows the association between the initial status and growth rate for each class. Figure 7 depicts how each subpopulation clusters with respect to the growth factors: PLPs have the lowest initial status with no growth, HPs have the highest growth rate with no growth, and LPs initial status is between the initial status of HPs and LPs with no growth. Finally Ss initial status is closer to PLPs with strong growth rate.

Figure 6. Reading Grades 3-6 HP, LP, S and PLP

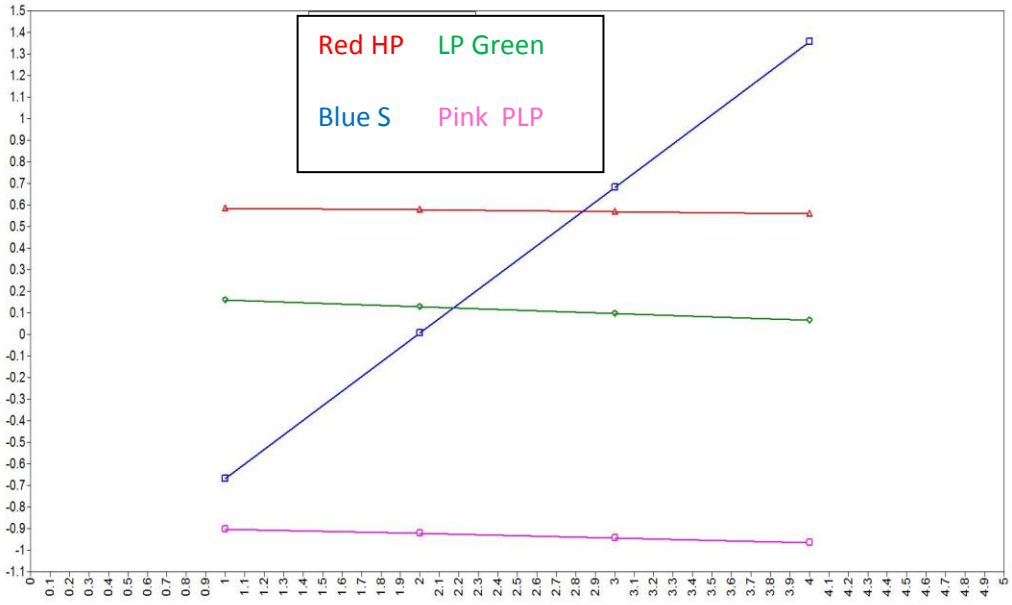
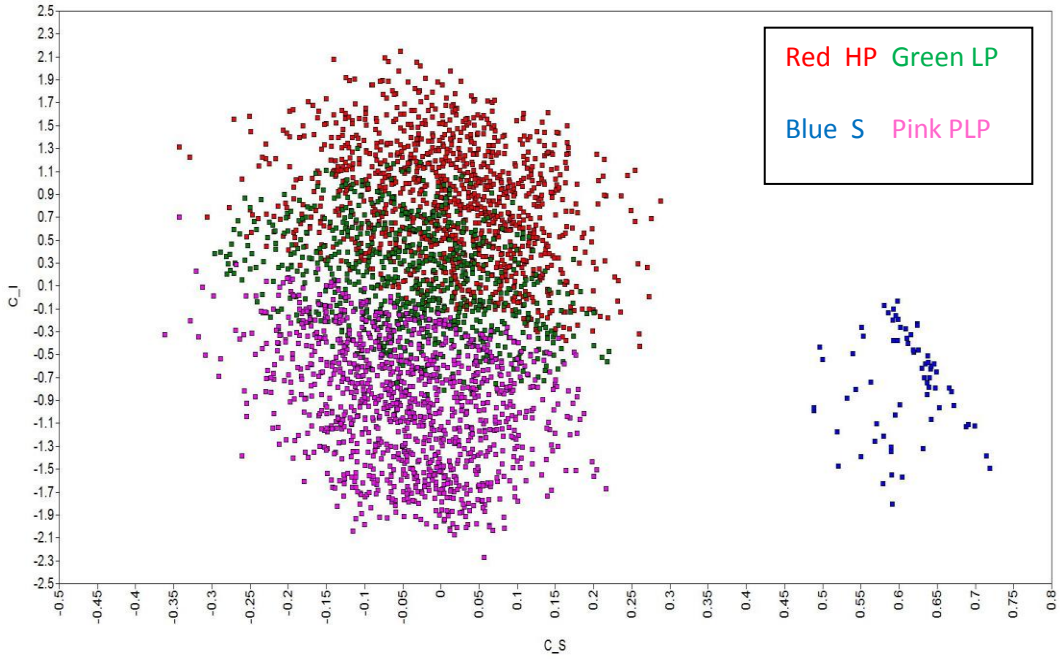


Figure 7. Reading Grades 3-6 Latent Classes Initial Status (C_I) and Growth Rate (C_S)



The within-level results (Table 23) indicate that the probability of membership in class 3 (S), relative to PLP (the poorly developing reference class 1), is statistically significant (1.123) for non LEPs (FRL has no influence). The probability of membership in class 2 (HP), relative to PLP, is statistically significant (0.798) for non-FRL students (LEP has no influence). The probability of membership in class 4 (LP), relative to PLP, is statistically significantly (0.728) for non-FRL students (LEP has no influence). The between-level results indicate that there is no statistically significant relationship between any developing classes and school SES (when PLP is the reference class).

Table 23. MLGMM Multinomial Estimates with School SES

Parameter	Estimate	S.E.	Est./S.E.
Within level			
S on			
LEP	1.123*	0.513	2.191
FRL	-0.333	0.633	-0.595
HP on			
LEP	0.474	0.43	1.101
FRL	0.798*	0.192	4.161
LP on			
LEP	0.09	0.25	0.362
FRL	0.728*	0.202	3.598
Between level			
S on			
SCHSES	1.246	0.819	1.521
HP on			
SCHSES	-0.928	0.508	-1.826
LP on			
SCHSES	-0.322	0.379	-0.849

Table 24 pertains to the within-level estimates and shows the results for the influence of both covariates on the growth factors intercepts for each class, residual variance for the growth factors for each class and residual covariance for the growth factors for each class. The influence of both covariates on initial status mean is positive and statistically significant (0.220 for LEP and 0.357 for FRL). Since I coded non-FRL students as 1, the non-FRL students had a higher initial status average in the reading scores at time 0 than did the FRL students . In addition, the influence of the FRL on growth rate was negative and not significant (0.012), which suggests that there was no difference with respect to growth rates between FRL and non-FRL students . Since I coded non-LEP students as 1, non-LEP students had a higher initial status average in the reading scores at time 0 than did the LEP students. The influence of LEP status on growth rate is negative and statistically significant (-0.031), which suggests (since I coded non-LEP students as 1) that non-LEP students have a more negative growth rate than LEP students. The residual variance for the

student initial status is 0.301 and significant, and the residual variance for the student growth rate is 0.018 and significant. The residual covariance between the growth factors is (-0.02) and significant, which indicates that students with higher initial status have lower growth rates or that students with lower initial status have higher growth rates. This result is similar to the one-class two-level MLM with school SES as level-2 covariate from section 4.2.2.

Table 24. MLGMM Within Estimates with School SES

		Estimate (Standard Error)			
Level	**Parameter**	**Class 1**	**Class 2**	**Class 3**	**Class 4**
	Within level				
Time	Residual Variances				
	Reading 1	0.139* (0.009)	0.107* (0.008)	0.622* (0.067)	0.029* (0.004)
	Reading 2	0.139* (0.009)	0.107* (0.008)	0.622* (0.067)	0.029* (0.004)
	Reading 3	0.139* (0.009)	0.107* (0.008)	0.622* (0.067)	0.029* (0.004)
	Reading 4	0.139* (0.009)	0.107* (0.008)	0.622* (0.067)	0.029* (0.004)
Student	Regression of Reading Intercept				
	LEP	0.220* (0.111)	0.220* (0.111)	0.220* (0.111)	0.220* (0.111)
	FRL	0.357* (0.072)	0.357* (0.072)	0.357* (0.072)	0.357* (0.072)
Student	Regression of Reading Growth Rate				
	LEP	-0.031* (0.011)	-0.031* (0.011)	-0.031* (0.011)	-0.031* (0.011)
	FRL	0.007 (0.013)	0.007 (0.013)	0.007 (0.013)	0.007 (0.013)
Student	Residual Variances for Reading				
	Initial Status	0.301* (0.028)	0.301* (0.028)	0.301* (0.028)	0.301* (0.028)
	Growth Rate	0.018* (0.001)	0.018* (0.001)	0.018* (0.001)	0.018* (0.001)
Student	Residual Covariance, reading initial status and growth rate	-0.020* (0.007)	-0.020* (0.007)	-0.020* (0.007)	-0.020* (0.007)

Table 25 pertains to the between-level estimates and gives results for the initial status *ib* intercept and growth rate *sb* intercept for the growth factors for each LC, the influence of school SES on the growth factors for each LC, the variance between the growth factors for each LC and the covariance between the growth factors for each LC. The four LCs are ordered from high to low reading achievement initial status intercept: 0.623 (class 2, 37%), 0.253 (class 4, 30%), -0.493 (class 3, 2%) and -0.718 (class 1, 31%). Here, the initial status intercept for classes 1 (PLP) and 2 (HP) are statistically significantly different from zero.

Table 25. MLGMM Between Estimates with School SES

		Estimate (Standard Error)			
Level	**Parameter**	**Class 1**	**Class 2**	**Class 3**	**Class 4**
School	Between level Regression of Reading Intercept				
	on School SES	-0.636* (0.118)	-0.636* (0.118)	-0.636* (0.118)	-0.636* (0.118)
School	Regression of Reading Growth Rate				
	on School SES	-0.034 (0.019)	-0.034 (0.019)	-0.034 (0.019)	-0.034 (0.019)
School	Intercept of Reading				
	for Initial Status	-0.718* (0.152)	0.623* (0.181)	-0.493 (0.254)	0.253 (0.161)
	for Growth Rate	0.020 (0.03)	0.037 (0.027)	0.722* (0.098)	0.012 (0.018)
School	Residual Variances for Reading				
	for Initial Status	0.019* (0.008)	0.019* (0.008)	0.019* (0.008)	0.019* (0.008)
	for Growth Rate	0.001* (0)	0.001* (0)	0.001* (0)	0.001* (0)
School	Residual Covariance, reading initial status and growth rate	-0.002 (0.001)	-0.002 (0.001)	-0.002 (0.001)	-0.002 (0.001)

Table 26 shows the results for each LC probability classification and Figure 7 shows the distribution of *ib* by trajectory classes. The growth rate intercept for these four classes are: 0.037 for class 2, 0.012 for class 4, 0.722 for class 3 and 0.02 for class 1. Here, only the growth rate intercept for class 3 (S) is statistically significant different from 0. The results for the classification of the latent trajectories are similar to the results of the previous sections (unconditional at level-2 MLGMM). With respect to the influence of school SES on the growth factors, increasing school SES lowers the school average reading performance at time zero (-0.636) whereas school SES does not influence a school's average growth rate (-0.034). The residual variance for the school initial status is 0.019 and significant, and the residual variance for the school growth rate is 0.001 and significant. The residual covariance between the growth factors indicates (when school SES is zero) that the two growth factors are not correlated so that reading initial performance mean in a school is not associated with school growth rate mean. This result is similar to the one-class two-level MLM with school SES as level-2 covariate from section 4.2.2.

Table 26. MLGMM Class Membership with School SES

Class	Initial Status	Growth Rate	Probability
Class 2 (HP)	0.623*	0.037	0.37
Class 4 (LP)	0.253	0.012	0.30
Class 3 (S)	-0.493	0.722*	0.02
Class 1 (PLP)	-0.718*	0.02	0.31

Table 27. MLGMM Fit with and without School SES

Model	Number of Classes	Number of Parameters	Log-likelihood	AIC	BIC	SABIC
Unconditional Level-2	4	31	-9849	19760	19950	19851
Conditional Level-2	4	36	-9816	19704	19924	19810

4.5 School Value-Added Estimates

In this section, I describe my systematic examination of the school value-added estimate differences (standard error and school value-added classification) using two types of methodologies (conventional MLM and MLGMM) by directly modeling the growth in a student's test scores over time. I estimate a set of four value-added scores: two with MLM (with and without school SES) and two with MLGMM (with and without school SES).

4.5.1 School Value-Added Formulation Specifications

Here I use reading scores from grades 3 through 6 (times or t = 0,1, 2 and 3), using sample sizes typically available in practice (1-5 years of data per school) to create the parameters for both frameworks. To calculate each school's value-added score I used the formula (observed score O_{oj} - predicted score Y_{4j}), where the observed score was the student's 2015 student reading score and the predicted score was the student's predicted estimate from each respective model framework (MLM and MLGMM) for t = 4, which represents students' gain score residuals for time 4 and school j. The overall average for the 2015 reading scores (which I standardized standardized to have an M of zero and an SD of 1, as I did for the 2011, 2012, 2013 and 2014 reading scores) was 454.047 with a S.D of 11.797. My study population was reduced by 5% to 3181 (95% of original population) because some students left the district in 2015.

Using the two model frameworks, I obtained two sets of school value-added score over five consecutive years, which I used to obtain the overall mean value-added and SDs across schools in the district; I later used the latter estimates to standardized the value-added scores using a *t* distribution. Table A1a (Appendix) shows the pre-standardized value-added score estimates for each school when school SES is not specified in the model and Table A1b shows the pre-standardized value-added scores estimates for each school when school SES is specified in the model. Figures A1a and A1b (Appendix) show a graphical depiction of the pre-standardized value-added score distribution for both methodological frameworks. The pre-standardized value-added estimates for the conventional MLM has heavier tails than the value-added estimates for the MLGMM when school SES is not specified. The MLM mean value-added estimate is -0.047 (SD 0.895) with a range of (-0.572, 0.725), while the MLGMM mean value-added estimate is -0.073 (SD 0.759) with a range of (-0.507, 0.549). When I specify school SES in the model, the pre-standardized value-added estimates for the conventional MLM also have heavier tails than the value-added estimates for the MLGMM framework (particularly for the schools with negative value-added scores). The mean value-added estimate for MLM is 0.031 (SD 0.864) with a range of (-0.482, 0.391), while the mean value-added estimate for MLGMM is 0.059 (SD 0.737) with a range of (-0.357, 0.424). For both scenarios using MLGMM narrows the distribution of the pre-standardized value-added scores.

Other parameters that can influence school value-added scores, including the school size (n) and the number of schools in the district (N), as well as the key variable of my analysis (school SES) also appear in Tables A1a and A1b (Appendix). According to the value-added literature (Yumoto, 2011, Schochet & Chiang, 2013) an effective sample size is equal to 32 students per school and all of the schools in my data exceed that number (Tables A1a and A1b).

4.5.2 Standardized Value-Added Scores and Thresholds Specifications

Any performance measurement system must contain a decision rule used to classify schools as meriting or not meriting special treatment. One of the most prevalent value-added models applied in practice is the EVAAS model used by the TAP (National Institute for Excellence in Teaching, 2009), which classifies each teacher/school into a performance category based on a *t*-statistic obtained by testing the null hypothesis that the school's performance is equal to the average performance in a reference group (see Solomon, White, Cohen, & Woo, 2007; Springer, Ballou, & Peng, 2008, Schochet & Chiang, 2013). Thus, hypothesis testing is an integral part of the policy landscape in performance measurement and forms the basis for my method of comparing school value-added estimates. In my formulation, students correspond to level-1 (indexed by *i*) and schools define level-2 (indexed by *j*), where school estimates of the standardized value-added scores are expressed in *t*-statistics scores (T_{ij}) for both models and I focus on these values in my school-level analysis.

In particular, I consider a performance measurement scheme that addresses the question, ''Which schools performed particularly well or poorly relative to the average school in the district using each methodological framework?'' I assume here a classical hypothesis testing strategy for both the MLM and MLGMM estimators. Under this scheme, my null hypothesis is $H_0 : T_j - \bar{\bar{T}}_{..} = 0$, where $\bar{\bar{T}}_{..} = \sum_{j=1}^{N} \frac{T_{.j}}{N}$ is the mean value across all schools in the district T_{ij} is the expected value-added score of a randomly chosen student *i* if assigned to school *j* and *N* is the number of schools in the district. Using this testing approach, I will identify for special treatment schools for which I reject the null hypothesis using a two-sided *t*-test. Under the EVAAS model, if a school's value-added estimate is found to be statistically significantly below or above average, district officials likely will then want to test whether the school's true performance is below or above average.

A critical issue of my testing approach is the determination of a threshold which defines a meaningful performance difference between schools (that is, the value of *T* in Figures A2a and A2b). Following the approach used elsewhere (Bloom, Hill, Black, & Lipsey, 2008; Kane, 2004; Schochet, 2008, and Schochet & Chiang, 2013), I identify an educationally meaningful threshold using the natural progression of student test scores over five years of time. I consider the threshold for this test to be t=2.021 with 40 degrees of freedom (calculated as the number of schools in the district minus 1) to achieve the conventional Type I error rate of α=0.05 with about 4 years of data. Under this approach adapted from that of the TAP I consider schools with *t* statistics (from standardized value-added scores) above 2.021 to have grown more than the district average (categorized as blue); when the school *T* score falls under -2.021 I consider the school red because it grew less than the district average, and the remainder of the schools I consider green because their growth is similar to the district average (National Institute for Excellence in Teaching, 2009).

Using this threshold criterion, when I did not include school SES at level-2 in the model, I identified eight schools as red, 11 as blue and 22 schools as green under the MLM methodology, and I identified 11 red schools, 12 blue and 18 green under the MLGMM framework. In contrast, when I specified school SES at level-2, I identified three schools as red, six as blue and 32 schools as green under the MLM methodology, and I identified five schools as red, 5 blue and 31 as green for the MLGMM framework. By adding school SES at level-2, the number of schools I identified for special treatment (red and blue schools) decreased by nearly half the original number for both methodologies (19 vs. 9 for MLM and 23 vs. 10 for MLGMM), while the number of schools identified for special treatment across methodologies is somewhat similar whether I specify school SES in the model or not (Tables A2a and A2b).

My main interest in this classification is to determine which schools change in performance status (e.g., from red to green) and in what direction when comparing the two methods. Reports from the literature indicate that performance systems utilizing conventional LM-based estimates schools' value-added yield a misclassification rate of 15% (Schochet & Chiang, 2013); in other words, individual using the system would erroneously identify or miss for recognition one seventh of schools. I found that seven schools (17%) changed status (e.g., from red to green) between the MLM methodology and the MLGMM methodology (Tables A2a and A2b, Appendix). As previously mentioned, I suggest that the conventional models do not account for important variability at the student and school level.

I expected that MLGMM in addition to school SES would be able to account for this variation within a diverse (as defined by the wide range in the proportion of students receiving FRL) urban district. Table A2a shows the standardized value-added school scores, school size and school SES when school SES is not included in the model. When school SES is not included in the model, the mean of the standardized MLM value-added scores distribution is 0.070 (SD 3.165) with a range of (-5.944, 5.557), while the mean of the standardized MLGMM value-added scores distribution is -0.082 (SD 3.233) with a range of (-5.702, 5.588). Figure A2a shows that the MLGMM standardized school value-added scores distribution has tails similar to those of the MLM. On the other hand, when school SES is specified in the model, the mean of the standardized MLM school value-added scores distribution is -0.085 (SD 1.692) with a range of (-3.742, 3.941), while the mean of the standardized MLGMM value-added scores distribution is -0.013 (SD 1.831) with a range of (-3.820, 5.174). Figure A2b shows that the MLGMM distribution has slightly more heavy tails than does the MLM particularly in the upper part of the distribution.

Taken together, the model without school SES at level-2 suggests that more schools are in need of special treatment than the model with school SES at level-2 for both methodological frameworks. The range of the school value-added scores are wider for the MLM (-5.944, 5.557) and for the MLGMM (-5.702,5.588) methodologies when school SES is not specified in the model than when the school SES is specified, (-3.742, 3.941) for MLM and (-3.820, 5.174) for MLGMM respectively (see Figure A7). When school SES is specified at level-2, 17% of the schools in the district change status with the more complex methodology, suggesting that they may have been erroneously classified. On the other hand, when school SES is not specified at level-2, there is only a 10% disagreement in classification between the methodologies.

4.5.3 School Value-Added Estimates ANCOVA Results

Here I will discuss a model that includes school SES. The first question I address is whether the average of the schools' standardized value-added scores obtained using the MLM methodology is equal to the average schools' standardized value-added scores calculated using the MLGMM methodology. I answered the question using a repeated measures analysis. I did not reject the hypothesis H_o: $u_{MLM} = u_{MLGMM}$ since I obtained F= 0.17 (p 0.683), using a Type I error rate of $\alpha = 0.05$, and thus I had no significant evidence to contradict the null hypothesis that the means of the schools' value-added between the two methods are equal.

Reports from the literature indicate some evidence for misclassification of schools with value-added scores in both extremes of the distribution, particularly when those schools are either very high SES schools or very low SES schools. I addressed whether school SES influenced the mean difference of the two methodologies. Using a repeated ANCOVA, I rejected the hypothesis H_o: $u_{MLM} - u_{MLGMM} = 0$ with a F=4.59 (p= 0.038), $\alpha = 0.05$; this suggests that the mean difference between methods was statistically different from zero, after controlling for school SES. I also rejected the hypothesis H_o: $\beta_{School\ SES} = 0$ was rejected with a F=6.64 (p=0.014),

$\alpha = 0.05$, which suggests that the mean of the standardized value-added scores were different by methodology when controlling for school SES. The estimated mean difference in the estimated effect of school SES was -1.288, i.e., for every unit increment of the proportion of students receiving FRL in a school, the difference between method means decreased by 1.29; for instance, for a school with only 13% of students receiving FRL the impact of school SES on average methodology difference is -0.17, however, for a school with 50% of students receiving FRL the impact of school SES on average methodology difference is -0.64, and for a school with 99% of students receiving FRL the impact of school SES on average methodology difference is -1.27. As a consequence, the gap of value-added scores between the methodologies is wider for homogenous schools, more specifically, schools with majority low (less than 21%) and majority high (99%) SES students. However, the gap of value-added scores is narrowed for schools with other compositions of low and high SES students (Table A3, Appendix). As expected, for schools with a very low percentage of students (less than 21%) receiving FRL, the MLM estimated value-added scores were larger than MLGMM's, and, the value-added scores were similar across methodologies for schools that were more heterogeneous in composition as defined by the percent of students receiving FRL (i.e., schools with 50 % of students receiving FRL and 50% of students not receiving FRL). I also expected that the school value-added scores for the MLM methodology would be larger than MLGMM's for schools with a very high percentage (about 99%) of students receiving FRL.

I ran similar tests when school SES was not specified in the model, and the results indicated no evidence that the means of the standardized value-added scores across methodologies were different (F= 0.60, p= 0.443). Figure A3a shows a graphic depiction of the standardized school value-added scores when value-added scores are ranked from low to high for the same school.

4.5.4 School Value-Added Precision Estimates

Given that individual schools can be subject to significant consequences on the basis of their value-added estimates, researchers have begun to pay more attention to the precision of these estimates. A number of studies have been conducted which examined the extent to which differences in single-year performance estimates across schools were due to persistent (or long-term) differences in performance—the types of differences I intended to measure—rather than to transitory student-level and school-level influences that induce random error, and thus imprecision, in the estimates (Yumoto, 2011; Schochet & Chiang, 2013).

I focused on grades 3-6 because there is empirical evidence available on key parameters affecting the precision of value-added estimates for those grades and pretests are likely to be available for analysis (EOGs start at grade 3). The precision of value-added estimates depends on several parameters such as the size of the school's, the number of schools in the district (N), the number of years used to model the predicted growth, the threshold of risk chosen and the methodological framework. The first two factors are a given because I am using empirical data from a district in North Carolina.

Having said that, the sample size per school is larger than the effective sample size of 32 recommended by scholars in this area, and the number of schools utilized is also aligned with the adequate number of schools used in previous practice and simulation studies (Yumoto, 2011; Schochet & Chiang, 2013). The third factor (the number of years in my models of the school predicted growth) is also adequate relative to models found in previous practice studies, about 5 years. that the precision of value-added estimates improves with the number of years included in the model; for instance, the reliability of the value-added estimator is .38 for 1 year, .65 for 3 years, .76 for 5 years, and .86 for 10 years.

Since my work is empirical, some of the factors were already defined (i.e., school sample size and number of schools in the district). I chose the values for other factors since previous work has indicated its adequacy (i.e., number of years to create the school value-added scores and risk threshold). Once a model includes all the factors mentioned above, any differences in the precision of the value-added estimates is expected to be due to the different methodologies used. I expected that using MLGMM would reduce the uncertainty of the value-added estimates compared to the MLM value-added estimates because MLGMM captures more of the relevant variability. One of my goals in this analysis was to show that student heterogeneity is a key source of imprecision in estimating differences in value-added scores across schools. On average, 92% of the total gain score variance is attributable to student differences (Schochet and Chiang, 2010; 2013), while the source of imprecision that stems from school-level factors accounts for, on average, 1% of the total variance in gain scores.

4.5.4.1 Descriptive Statistics

Table A1a (Appendix) shows the pre-standardized value-added score SDs for each school in the data when school SES is not specified. When school SES is not specified in the model, the mean SD of school value-added scores for MLM is 0.838 (SD 0.095) with a range of (0.650, 1.056), while the mean SD of school value-added scores for MLGMM is 0.709 (SD 0.114) with a range of (0.480, 0.952). Figure A4a (Appendix) shows a graphical depiction of the distribution of the school value-added scores SDs for both methodological frameworks. As the graphic shows the distribution of the school value-added SDs for the conventional MLM is shifted to the right of the distribution of the school value-added SDs of the MLGMM framework and the overall average value-added scores SDs across schools tend to be larger in magnitude for the MLM framework than for the MLGMM's.

Table A1b (Appendix) shows the pre-standardized value-added score SDs for each school in the sample when school SES is specified. The mean of the value-added SDs for MLM is 0.838 (SD 0.096) with a range of (0.649, 1.070), while the mean value-added SD for MLGMM is 0.717 (SD 0.123) with a range of (0.483, 0.986). Figure A4b (Appendix) shows a graphical depiction of the school value-added scores distribution for both methodological frameworks. As the graphic shows, the distribution of the school value-added SDs for the conventional MLM is shifted to the right of the distribution of the school value-added SDs of the MLGMM framework. The overall average value-added scores SDs across schools tend to be larger in magnitude for the MLM framework than for the MLGMM's. In summary, the MLGMM framework appears to yield smaller SDs than the MLM framework regardless of whether the model specifies school SES.

Table A2a (Appendix) shows the standardized value-added score standard errors for each school in the data when school SES is not specified. Figure A5a (Appendix) shows a graphical depiction of the value-added scores distribution for both methodological frameworks. As Figure 5a shows the distribution of the MLM school value-added standard errors is shifted to the right of the MLGMM's distribution. The mean of the school standardized MLM standard errors distribution is 0.099 (SD 0.018) with a range of (0.066, 0.146), while the mean of the school standardized MLGMM standard errors distribution is 0.084 (SD 0.020) with a range of (0.056, 0.144).

Table A2b (Appendix) shows the standardized value-added score standard errors for each school in the sample when school SES is specified. Figure A5b (Appendix) shows a graphical depiction of the value-added scores standard errors distribution for both methodological frameworks. As Figure 5b shows the distribution of the MLM school value-added standard errors is shifted to the right of the MLGMM's distribution. The mean of the school standardized MLM standard errors distribution is 0.099 (SD 0.018) with a range of (0.066, 0.146), while the mean of

the school standardized MLGMM standard errors distribution is 0.085 (SD 0.022) with a range of (0.056, 0.154). To summarize, the overall average value-added scores standard errors across schools tend to be larger in magnitude for the MLM framework than the MLGMM's regardless of whether school SES is specified in the model (Figure A8, Appendix).

4.5.4.2 ANCOVA Results

In this section, I first present results that pertain to schools when the variable school SES is specified in the model. The first question I answer is whether the average of the schools' value-added standard errors for the MLM methodology is equal to the schools' average value-added standard errors for the MLGMM methodology (after the value-added scores from both methodologies were converted to natural logarithm). I addressed this question using repeated measures analysis. I rejected the hypothesis H_o: $SE_{MLMmean} = SE_{MLGMMmean}$ with a F= 45.65 and $p < 0.0001$, using $\alpha = 0.05$, suggesting that the means of the schools' value-added standard errors between the two methods were different ($Mean\ Difference_{MLM-MLGMM}$ = 0.014). Figure A6b (Appendix) shows the standard errors for the same school for both methodologies ranked from low to high values by the MLM framework; as the graphic shows, the MLM framework consistently yielded higher values for the standard errors than did the MLGMM framework. The exceptions were two schools (26 and 7) with sample sizes of 36 and 35 students, respectively. These schools had 99% and 97% of students in FRL respectively; one school was in the green category (i.e., grew their students about to the same as the district growth average, therefore it met the district growth standards) and the other was in the red category (i.e., the school grew their students below the district growth average, therefore the school did not meet district standards) by both methodologies respectively. I obtained similar results when school SES was not specified in the model (F=59.70, p <0.0001 and $Mean\ Difference_{MLM-MLGMM} = 0.015$). Figure A6a (Appendix) shows the standard errors for the same school for both methodologies ranked from

low to high values by the MLM framework; as the graphic shows, the MLM framework consistently yielded higher values for the standard errors than did the MLGMM framework. The exceptions were four schools (34, 24, 1 and 26) with sample sizes of 79, 106, 51 and 36 students respectively. These schools had 88%, 98%, 91% and 99% of students in FRL, respectively; both methodologies indicated that the first two schools were in the green category and the other two as blue, respectively. Overall, the average school standard errors for the MLM are larger than for the MLGMM regardless of whether school SES is specified in the model or not.

In addition to the previous analysis, I used a repeated measures ANCOVA to assess if the sample size of the schools(n) might influence the average difference of the standard errors between both methodologies. However, I did not find that it did, since I obtained (F = 1.33 and p = 0.26). The standard errors for the MLM are larger than for the MLGMM methodology regardless of the sample size of the schools. I obtained similar results when school SES was not specified in the model (F= 0.22, p= 0.64).

I also found that the quality of estimates was better for the MLGMM framework based on the ratio of the size of the standard errors taken into consideration with regards to the value-added estimates. For instance, when the standard error is equal to or larger than 50% of the value-added estimate, the value-added estimate may be considered questionable (Corcoran, 2010). When school SES is present in the model, I found that the MLM framework yielded six schools (37, 35, 36, 32, 6 and 14) with questionable value-added scores based on this ratio, while the MLGMM yielded only three schools (14, 41 and 33) with questionable value-added estimates. The schools identified as questionable by MLM (five of which were indicated as green category by both methodologies) have sample sizes of 116, 116, 41, 112, 96 and 55, respectively, with 51%, 46%, 89%, 48%, 19%, and 94% of students in FRL. School 6 was indicated as green category by MLM and red category by MLGMM. The schools identified as questionable by MLGMM have sample

sizes of 55, 49 and 57 (respectively) with 94%, 15%, and 17% of students in FRL. All these schools were indicated to be green schools by both methodologies.

When school SES is not present in the model, I found that the MLM framework yielded two schools (12 and 31) with questionable value-added scores, while the MLGMM yielded only 1 school (18) with questionable value-added estimates. The schools identified as questionable by MLM have sample sizes of 67 and 82 (respectively) with 92% and 51% of students in FRL. All these schools were indicated to be green schools by both frameworks. The school identified as questionable by MLGMM has a sample size of 118 with 59% of students in FRL. This school was indicated to be green by both methods.

4.5.5 School Value-Added Classification Disagreement Rates

Another of my goals was to show that student heterogeneity is the key source of misclassification of schools, thus, I next focus on potential false positive (false discovery) and false negative error (false non-discovery) rates to measure the accuracy of a performance measurement system based on hypothesis testing and methodological framework type (MLM vs. MLGMM). Under the EVAAS scheme, type I error is the probability of recognizing a school for special treatment (or being statistically significantly different from the district average) when it is not, while Type II error is the probability of failing to recognize a school for special treatment when it should be.

To define the threshold indicating sufficient evidence to conclude that a school requires special treatment, I set threshold of risk at five percent (Schochet & Chiang, 2013), thus I would reject the null hypothesis if the probability of observing a value as extreme or more extreme is less than or equal to 0.05. If the probability of observing a value as extreme or more extreme is greater than five percent, I will not reject the null hypothesis. I found that when school SES is specified in the model, the school value-added classification yielded a disagreement rate between

methodologies of approximately 17% (a total of seven schools, Table A5b, Appendix) using a threshold of risk of 5%. Based on my Kappa test results, I concluded that there was no sufficient evidence for agreement between the methodologies when school SES was specified in the model ($|Z|= 2.69$, $p=0.22$). MLGMM indicated that three schools were red (but MLM indicated that they were green); MLGMM also indicated one school as a blue school but MLM indicated that it was green. Assuming that MLGMM represents the true model, those using the traditional model would have missed for recognition 10% of high and low performing schools in the district.

On the other hand, they also would have erroneously identified for recognition seven percent (three schools, see Table A5b) of persistently average schools (MLGMM indicated three schools but MLM indicated that one was red and two were blue). Meanwhile when school SES was not specified, I found a disagreement rate between methodologies of about a 10% (a total of four schools, Table A5a) using a threshold of risk of five percent. Again assuming that MLGMM represents the true model, those using the conventional model results would have missed for recognition 10% (a total of four schools) of high and low performing schools with MLM. In addition, I concluded based on my Kappa test results that there was some evidence for agreement between the methodologies when school SES was not specified in the model ($|Z|= 4$, $p < 0.0001$). Taken together, methodology type impacts disagreement rates, but only when school SES was specified in the model.

I also explored the disagreement rates if I used a less restrictive threshold of risk (10%) or if I used a more restrictive threshold of risk (one percent), compared to the threshold of risk at 5%. My main interest in this classification is to assess which schools change in performance status and in what direction based on changes in the criterion threshold so that policymakers can find an acceptable criterion which indicates a high or low performing school when using value-added estimates so that they may make their high-stakes decisions regarding schools using the

best approach. I next describe the disagreement rates of the full model with school SES at level-2 when the threshold of risk is at 10% and at one percent. My results indicate that the disagreement rate increased by about eight percentage points (25% from 17% with threshold risk at five percent) using a threshold of risk of 10%, and that it decreased by about 10 percentage points (7% from 17% with threshold risk at 5%) using a threshold of risk of one percent (Table A5b). Assuming that MLGMM represents the true model, when the threshold of risk is at 10%, policymakers would have failed to identify about 10% (a total of four schools) of schools for special treatment and would have falsely recognized for special treatment 15% (a total of six schools). Using the one percent threshold of risk, policymakers would have falsely identified about five percent (a total of two schools) of schools for special treatment and would have missed identifying about 2% (one school) of the schools for recognition.

In summary, when the threshold risk value becomes more liberal (10%), the disagreement rate increases. However, when the threshold risk value becomes more stringent (1%), the disagreement rate decreases. I observed similar patterns when school SES is not specified in the model but the disagreement rate was less dramatic (Table A5a).

CHAPTER V

DISCUSSION

In this chapter I discuss how my results in Chapter 4 address my main research questions. As described in Chapter 3, my goal was to investigate the impact on the school's value-added effect classification and its precision estimates resulting from accounting or not accounting for the subpopulations that might exist at the student-level (level-1) of analysis, as well as the impact of accounting for different proportions of students receiving FRL (as a proxy for school resources) within a single school (level-2).

Since schools do not have easy access to variables that I consider to be relevant since they represent factors known to influence student learning (parental involvement, private tutoring and others), they use demographic variables as proxy variables to account for the key variables they cannot access. However, the use of demographic variables is problematic because the fairness of the evaluation cannot be established if there was systematic bias resulting from not including key variables in the estimates of any school's value-added score.

For this reason, I set and defined a variety of conditions to investigate the magnitude of classification changes in schools' value-added effect estimates resulting from heterogeneous student growth within a school at level-1 (student-level) and to determine whether school SES at level-2 captures the key variability of each school potentially also affecting estimates of school value-added. I tested the use of four conditions: two MLM schemes with and without school SES and two MLGMM schemes with and without school SES. Prior to these tests I systematically tested a series of other conditions to established the number of LCs that best fit the MLGMM framework and I also tested the systematic inclusion of manifest variables (FRL and LEP) at

student-level to assess their contributions and interpretation in both modeling frameworks. My research questions were:

1) Are the classification results and precision of value-added scores estimates in the MLM affected by not accounting for LCs (potentially incorrectly-modeled level-1 effects)?
2) Are the classification results and precision of value-added scores estimates in the MLM affected by not accounting for LCs (potentially incorrectly-modeled level-1 effects) and by not accounting for school SES at level-2?

As I mentioned previously, I found that:

Value-added score estimate changes in MLM frameworks result in systematic changes in school value-added classifications, especially when the variable school SES is added to the model. The changes in classification diminished when school SES is not specified at level-2 in model. When school SES is present at level-2, the classification disagreement between methodology frameworks increased to 17%, especially for schools with a very high (>91%) or a very low proportion (<30%) of students receiving FRL. In addition, potential level-1 and level-2 model misspecification results in systematic changes in the magnitude of school value-added scores estimates, especially for schools with less variability in with respect to the proportion of students receiving FRL; these changes increased (>|0.5|) for schools with a very low proportion of students receiving FRL (<21%) and (to some extent) when the proportion of students receiving FRL is very high (>91%). The magnitude of the difference across methodologies is a bit smaller for poorer schools than for richer schools.

I conclude from this that the larger differences between these methodology frameworks were not located in the extreme of the school value-added scores distributions as I had hypothesized, however, there were differences between methodologies particularly for the richest schools and to a lesser extent for the poorest schools when school SES is specified. In addition, I

found that the school value-added scores estimates, when MLM is used, are more potentially bias in favor of the rich schools and potentially biased against the poorest schools when school SES is specified. These results are similar to those found in Yumoto's (2011) simulations. Although this study cannot ascertain for sure which model was best as Yumotos's did (because his study was based on simulations and therefore he knew which model was the true model), I infer that accounting for level-2 and level-1 heterogeneity provides the better specified model (because it accounts for relevant systematic variability). In addition, there is some evidence that indicates that MLGMM fits the data better than MLM but I will discuss this point in more detail in later sections.

I also found that the distribution of school value-added scores was considerably less heavy tailed when school SES was accounted for at level-2 regardless of the methodological framework I used; thus I categorized significantly fewer schools as needing special treatment (a reduction of more than half in red schools or and about half in blue schools). Methodology framework had much less of an impact on the extension of the school value-added distribution tails than I had hypothesized it would except for a few conditions described Section 5.3. The MLM methodological framework produced a heavier tail than did the MLGMM only for the lower part of the distribution or for schools with negative value-added scores when school SES was not specified in the model. However, when school SES was not specified at level-2, I categorized about a third more schools as red with MLGMM than with MLM and about one tenth more schools were classified as blue with MLGMM than with MLM; when school SES was specified at level-2, I categorized about half more schools as red with MLGMM than with MLM, and about one seventh less schools as blue with MLGMM than with MLM. I conclude from this that specifying school SES at level-2 resulted in the largest impact on school value-added classification at the tails of the distribution.

With regards to the value-added scores precision estimates, I found that the methodology I used resulted in systematic impacts on the precision (as measured by standard errors) of the estimated school value-added effects; however, unlike its effects on the value-added scores, the specification of school SES at level-2 did not impact precision estimates. The effects of methodology framework on precision tended to be consistent whether school SES was specified in the model. Standard errors were smaller on average for the MLGMM framework than for the MLM framework.

Taken together, these results suggest that the evaluation of schools, in terms of effects (effectiveness), using VAM, can potentially change in different contexts (methodology, school variability accounted by school SES at level-2 and the interaction of both). Using the MLGMM methodology enabled me to control these sources of systematic variation; those not using this methodology may obtain potentially biased results that could lead to a greater degree of (misplaced) confidence in such potentially incorrect estimates.

More disturbing is the *pattern* of results – a positive school value-added score estimates- for schools with low proportion of student receiving FRL and a negative school value-added score estimates- for schools with high proportion of students receiving FRL - for school effectiveness. My results suggest that schools will appear to be increasingly "better" due to the MLM-potentially engendered bias and overestimation of positive cluster (school) effects, and that other schools will appear to be increasingly "worse" due to a similar overestimation (bias) of negative school effects.

In sections 5.1 and 5.2 I discuss the results on how I systematically build the models for both methodological frameworks. I introduce the results of this process because I consider the models' specifications as part of the discussion of the methodological frameworks quality. Particularly, I find these results relevant because they provide some evidence in which I can infer

which model is best. There are differences in the value-added scores estimates but in addition to those results, I also search for evidence that indicates which model is potentially less biased and more precise.

5.1 Model Identification Process with MLM

In this particular set of analyses, I address the following research questions:

1) Does the model fit improve when adding any of the manifest variables at level-1 and/or at level-2?
2) Does the interpretation of the manifest variables at level-1 and level-2 change by adding systematically each covariate (first FRL at level-1, then LEP still at level-1, continue by adding cluster (i.e., school) level of analysis and finally by adding school SES at level-2?

I briefly summarize my findings in this area: Adding the manifest student-level covariates (FRL and LEP at level-1) systematically did not improve the fit of the model for the one-level MLM and the two-level MLM model nor did adding the manifest covariate at level-2, school SES. Overall, the average initial status was negative and statistically significantly different from zero for all conditions with covariates at level-1 and when school-level is added; the initial status was positive but not statistically significant when school SES was accounted for at level-2. The growth rate was not statistically significant different from zero for all five conditions. I found a negative statistically significant strong effect of FRL on initial status but I did not find an impact of FRL on growth rate across all pertinent conditions . Also, I found a negative statistically significant moderate effect of LEP on initial status but I did not find an impact of LEP on growth rate across all pertinent conditions. Taken together, these results suggest that the interpretation of the manifest variables at level-1 do not change by adding systematically each covariate (first FRL at level-1, then LEP still at level-1, continue by adding cluster (i.e., school) level of analysis and finally by adding school SES at level-2).

5.2 Model Identification Process with MLGMM

In this particular set of analyses I discuss my modeling process when accounting for LCs at level-1.

5.2.1 Information Criteria Performance

I used six information criteria (BIC, SABIC, BICB, AIC, AICc and AIC3) to identify the model that fit the data best among the 1-, 2-, 3- and 4-class MLGMMs. Overall, all six of these information criteria performed fairly similarly across each of all modeling conditions, agreeing on the best model. The fit indices similar results performance in this study could be due to the fact that schools in the district for the most part have an adequate sample size of at least 40 students (with a few exceptions) and the district has an adequate number of schools, resulting in an overall adequate sample size. In this section I address the following research questions:

1) Which model fit the data best among the 1-, 2-, 3- and 4-class MLGMMs according to these six information criteria?
2) Do the growth trajectories class membership change by adding systematically (across the five conditions) the manifest variables at level-1 and by adding the multilevel analysis with the manifest variable at level-2?
3) Does the fit of the optimal class model was improved as manifest variables were systematically added to the model at level-1 and level-2 (across the five conditions)?

I briefly summarize my findings in this area: The best model fit had 4- classes (which closely aligned with the four profiles of H, PLP, S, and LP) when the model was unconditional. We will see in limitations that I centered each year scores to a mean of zero assuming an average linear growth rate, however each latent class can have its own growth trajectory (linear or not). HPs had the highest initial status with zero growth rate, PLPs had the lowest initial status with very small negative growth, the S group had a much lower initial status than expected but a

positive and strong growth rate, and LPs had a initial status between the HP and PLP with a very small negative low growth rate. The class membership did not change by adding systematically (across the five conditions) the manifest variables at level-1 and by adding the multilevel analysis with the manifest variable at level-2. The basic rankings of the four growth trajectories initial status did not change across the five conditions when I systematically added the manifest variables at level-1 or when I added the multilevel analysis with the manifest variable at level-2, but the growth rate for LPs and PLPs became null, while the growth rate for High Performers and Strivers remain the same. The fit for the four class model was improved (in terms of relative performance) as I systematically added variables to the model at level-1 and level-2 (across the five conditions), particularly when FRL was added to the model.

5.2.2 MLGMM Convergence and Interpretation of Effects

Combining multilevel structure with GMM did not seem to affect model convergence because although the cluster size varied across all schools, all clusters (i.e., schools) maintained the minimal effective size. In this section I discuss the indirect effect of FRL, LEP and school SES on initial status and growth rate through its influence on the classification of classes and address the following research questions:

1) Does the interpretation of the manifest variables at level-1 (FRL and LEP) and level-2 (school SES) change by adding systematically each covariate? Does the interpretation of the manifest variables at level-1 and level-2 change with the MLGMM framework from the MLM framework?

I briefly summarize my findings in this area: The MLGMM with 4- classes yielded a positive moderate higher significant initial status for high SES students compared to low SES students but there was no difference in growth rate between the two groups. These effects were consistent across all five conditions of the MLGMM framework modeling. The MLGMM

growth rate was comparable with what was found for the single class MLM but the effects were more attenuated. The MLGMM with 4- classes yielded a small positive higher significant initial status for non-LEP students compared to LEP students. However, non-LEPs had a very small negative significant growth rate compared to LEPs. The difference in initial status between LEPs and non-LEPs were smaller for the MLGMM framework than for the MLM framework but the growth rate difference became negative in the MLGMM framework (from zero in the MLM framework) . The single class MLM showed no difference between the growth rate of LEPs and non-LEPs, but MLGMM effects yielded a small but negative growth rate for non-LEPs compared to LEPs. Also, level-2 covariate school SES had a negative significant strong effect on initial status, but no statistically significant effect on growth rate. However, the impact of school SES on initial status was a bit stronger for the MLM framework than for the MLGMM framework with similar no statistically significant effect on growth rate across methodologies.

5.3 School Value-Added Effects

In this section I discuss the extent of differences in value-added scores in classification and precision by methodological framework.

5.3.1 Results on School Value-Added Classification

My main interest in this classification was to assess which schools changed performance status category (i.e., red to green) and in what direction when I used MLGMM instead of MLM and when I added school SES to the model to account for variability at level-2. I briefly summarize my findings in this area: Adding school SES to the model greatly affected the number of schools designated for special treatment (blue or red schools) more so than model methodology. By adding school SES at level-2, the number of schools identified for special treatment decreased by about half for both methodologies. I infer that school SES greatly

decreased school classification for special treatment because school SES accounted for the lack of random assignment from students to schools. VAMs assume that students are randomly assigned to schools but in reality schools are embedded in neighborhoods with a given SES. As a result, schools in poor neighborhoods will have a high proportion of poor students and schools in rich neighborhoods will have a high proportion of rich students. When school SES is added to the model, it accounts for the level-2 variability that potentially causes systematic bias into the value added scores estimates created by the large number of students with very low or very high initial status within a school.

The dramatic change in classification when adding school SES provides some evidence of the relevance of this variable as a source of systematic variation. I did not find such a dramatic change in classification when LC was added at level-1, as I expected, because the manifest variables may have been sufficient to capture level-1 variability. Particularly, FRL, as evidenced by the improvement of model fit with MLGMM. It is also worth mention that this model improvement with FRL was only observed when LC was present in the model at level-1 (with MLGMM), and that the conventional model did not captured this improvement. Because adding latent classes to the student level creates a blocking effect in which students of the same trajectory class are compared with regards to SES creating a finer grained analysis. In addition, methodology framework also affected which schools I categorized as needing special treatment but to a lesser extent than my addition of school SES to the model. I conclude that is the combination effect of MLGMM with school SES that has the greatest impact on school value-added scores classification.

5.3.2 Results on School Value-Added Magnitude Effects

I found some evidence of the effects of school SES on the mean school value-added scores by methodology framework. As expected, for high SES schools as well as for low SES

schools, the magnitude of the MLM absolute value-added scores was greater than those produced by MLGMM. Further, the MLM value-added scores were larger for high SES schools and smaller if not negative for low SES schools; I did not observe this when I used MLGMM.

I found in my results a sensitivity similar to that reported by Yumoto (2011) for the effects of school variability with respect to the mixture proportion of students in a given SES per school, which suggests that a higher proportion of high or low SES students within any school will create more extreme value-added scores when using the traditional framework rather than MLGMM. This further suggests that student heterogeneity contributes to the inflation of the school value-added score estimates; in other words, the greatest amount of potential bias would occur in a school with the smallest proportion (or the highest proportion) of low SES students within a school district that has a well distributed range of mixture proportion of high and low SES schools (i.e., a school district with high variability). Consequently, all schools, good or bad, would be most affected if evaluated in the context of heterogeneous districts (such as the one representing this study).

5.3.3 Precision of Estimates

As I expected, the precision of my estimates schools' value-added scores produced using the MLM framework tended to be worse regardless of whether school SES was specified in the model, which only serves to compound the problems associated with misclassification of the model. Also as I expected, the precision of estimates was not impacted by the number of students in each school.

5.4 Limitations of the Research

My work has several limitations. The first of these is the generalizability of my results – I examined only one district with its own distinctive school specifications (i.e., rural vs. urban

district, school district size and others). The composition of school SES for this district might not reflect reality in other more homogenous districts such as rural districts. I did not include LCs to represent the cluster level (e.g., between-level or school's level) where interactions between individuals and schools are very likely. The impact of a sizeable proportion of fast growth group members was especially apparent even in very low SES schools and this I did not expect (Table *A4*). The use of level-2 LCs to represent the different growth groups could potentially accentuate the differences found when schools SES was very high by methodology framework, and could more clearly demonstrate the differences in inferences that are supported by the MLGMM and MLM model conditions.

Another challenge was the standardization of the reading scores to a mean of zero and SD of 1. When I converted the scores, time score 3 yielded negative variance (Heywood case) and with that I had to constrain these variances to be equal for each time score across time. I carried over that same restriction when I use MLGMM, resulting in each LC having their own variance but equal across time as well. This procedure restricted the number of parameters that could be free. More specifically, I was limited to equal all parameters for the individual-level variables (FRL and LEP) across LCs. If I would not have had to contend with a Heywood case, I could have obtained richer information for each of these parameters at the individual-level for each LC. This would have allowed for a finer grained understanding of the contribution of each variable within each growth trajectory and, ultimately to start problem solving about specific solutions for this subgroups.

Another challenge was the issue of inferential robustness, i.e., when alternative models with a similar model fit (i.e., information criteria select alternative models) may lead to different interpretations or conclusions, which I did not fully been address. The conditions I specified (i.e., assessing 2 methodological framework, up to four-classes with MLGMM and school SES

specification at level-2) were intended to minimize the influence of uncontrolled effects to avoid this issue. However, in more complex real life data, it is extremely important to carefully investigate alternative models in order to make a valid interpretation of results, including the possibility of adding more LCs to the model if the data permits this.

Finally, I could not assess the proficiency of the schools in the district but only that schools with more positive and more negative cluster effects will actually generate potentially differentially biased estimates.

5.5 Future Directions

A question remains regarding how my results may be utilized to substantiate the use of value-added scores in the evaluation of schools, given that the end of year assessments were not developed for school accountability, particularly for the LEP population. In this section, some suggestions are discussed to improve the sound interpretation and use of value-added scores as an accountability index that supports student learning.

The intended interpretations and use of EOGs was to measure whether students learned the curriculum for their given school grade. EOGs also inform stakeholders whether students met school curriculum expectations and if not what potential remedies can be implemented to solve this issue (Moss, 2016). Drawing on Kane's work, this is what he called the direct use of test scores. The use of test scores rely directly on the information scores provide about measured constructs, including instructional guidance, student placement, comparisons among educational approaches, and educational management (Haertel, 2013).

On the other hand, Kane also alludes to the indirect test uses as mechanisms of action, leading to intended or unintended consequences, that do not depend directly on test scores (Haertel, 2013). Value-added score use is an example of indirect use and one of its purpose is to serve as a tool to sanction schools, teachers and principals if students grow below a given

expected value. The indirect use of test scores and its potential unintended consequences strengthens the argument for the importance of validity inquiry for these uses of assessment. For instance, as it was described in this study, by changing just a couple of variables (i.e., including LC at level-1 and school SES at level-2), different schools were classified for special treatment, particularly schools with majority low and high SE students. In one model specification (i.e., model without school SES at level-2), half of the schools were classified for special treatment and with the other model specification (i.e., model with school SES at level-2) only twenty percent of schools were selected for special treatment.

The schools with majority low SES students were more negatively affected (i.e., because they received worse classifications) and the schools with majority high SES students were positively affected (i.e., they received better classifications). This instability of value-added scores classification should concern us particularly when value-added scores are used to make decisions regarding different stakeholders (teachers, principals, students or schools) and that some stakeholders under some circumstances could be more negatively affected (i.e., schools with majority poor students).

In this study, the most troubling result is that the model specification that does not control for school poverty at the cluster-level classifies a larger number of the poorer schools to be sanctioned (when they should not be according to the model that controls for school poverty). For this very reason, unintended consequences should be carefully considered when making decisions about specific stakeholders, in this case schools. For instance, Kane (2006, 2013) orients his discussion of test use on a 'decision rule, which stipulates that certain actions be taken given certain test scores' (p. 46). He argues that decision inferences must be evaluated in terms of their consequences: they 'require evidence that the procedure achieves its goals without unacceptable negative consequences' (p. 15). For this particular study, the conclusion would be to

use the classification that causes damage (i.e., in terms of schools being sanctioned with funding restrictions or other type of punishment) to the least number of schools in the district until we have more robust evidence on which model specification inferred the most accurate value-added scores. For accountability purposes, it is relevant to consider the reliability, precision and consistency of the value-added scores' estimates as progress indicators. However, current research has not provided optimistic support. Research shows that value-added scores' classifications change based on several factors such as models used, variables specified, constructs used, sample size and others. Finding evidence of the value-added scores' estimates consistency is paramount in evaluating the validity of test-based accountability systems (Braun, 2016).

Current validity theory has provided explicit guidance about uses – decisions and actions, however, the focus has remained on intended decisions and actions associated with test scores (Cronbach, 1988; Haertel, 2013; Kane, 2006, 2013; Messick, 1989; Shepard, 1993). However, considering the test as the primary source of evidence under anticipates the complexity of how test scores are being used locally, in practice, by teachers and other education professionals in different contexts for their own purposes (Moss, 2013). Spillane (2012) explored relationships between data and how school officials used data to make decisions. Spillane (2012) argued that the way practitioners noticed and interpreted information is not based on research only but on other types of information guided by the organizations they are embedded in, such as local leadership type, their own professional point of view (or lack of) or other institutional norms and culture. In summary, Spillane (2012) argues that data practice should be framed in terms of several aspects of the organization and he advocates more research on the quality of data use in education, especially test data in an effort to get more and better data-based decision making in schools.

Future work in this domain should explore research that produces more robust evidence whether value-added scores classifications are a valid metric for school accountability because they are being used by school stakeholders (i.e., board members) without the pertinent statistical expertise and who are constrained by different types of pressures (i.e., whichever is the political agenda maybe) that have nothing to do with student learning. Following Chalhoub-Deville's (2016) argument, maybe the proof of validity should fall on the shoulders of the researchers who produced those scores given the gravity of the decisions being made, and these researchers should share the responsibility of unintended negative consequences with the school officials who make those actual decisions because it is their responsibility to have a sound strategy for student growth. With respect to accountability, the interpretations based on value-added scores can lead to rewards or sanctions for teachers, schools and districts. For instance, these rewards and sanctions may influence the number and type of teachers who teach in certain schools. In this school district, the current accountability index (based mainly on student Percent Proficient) has a correlation of -0.75 with school SES. This indicates that schools with a higher proportion of students receiving FRL tend to have a worse school classification (D or F). However, the correlation of my best model (MLGMM with school SES at level-2) with school SES is only 0.21. In this particular case, the state accountability index creates a negative impact for very poor schools because it does not take into consideration the initial status of the students. The current state accountability metric demoralizes teachers because this accountability index does not measure progress. As evidence of that, schools with negative grades tend to have over a 20% teacher turnover rate per year on average while schools with good grades tend to have teacher turnover rates of less than 10% (for this district). It is not surprising that with the existing accountability metric, schools with majority poor students, tend to have a bad grade but in addition to that, poor schools also lose a good proportion of the teachers every year creating an

internal managerial crisis within the schools. Principals must invest time and resources to find new teachers (who would prefer to work in a high SES school because they know student progress is not relevant) instead of utilizing that time to develop strategies of student growth. Using the adequate accountability metric index to evaluate schools or teachers should encourage teachers to improve student learning. The degree to which this goal is not realized and other damaging unintended consequences are not minimized is essential to investigating whether the appropriate metric is being used for accountability purpose.

Within this line of action about validity research on unintended consequences, further research is needed on whether there is agreement between school value-added scores with another subject matter (i.e., math) and the reading construct I used. Agreement in value-added scores' classifications between different constructs could provide stronger evidence of the validity of the use of VAM for school accountability. In the context of accountability, the degree to which accountability indicators are congruent with other indicators of school effectiveness can provide valuable supporting evidence for aggregate indicators such EVAAS. States like Minnesota and Colorado use a growth model as an accountability index in which they look at multiple indicators based on three constructs: math, reading and English language proficiency. In addition, the state of Minnesota makes a comprehensive investigation of the relationships between the accountability indicators, mentioned above, with other indicators of school effectiveness such as graduation rates and school attendance. This process of gathering evidence based on the relationships of accountability indices and other variables can be helpful to evaluate accountability measures based on students' test performance .

To extend the evidence that supports the use of value-added scores as a valid accountability index , I also suggest MLGMM-based scores should be tested using a greater range of real-world conditions. The introduction of between-level LCs to incorporate, or explore,

interactions between the between-level (e.g., schools) and the within-level (e.g., students) LCs might be useful. Although school SES is an important variable, school SES is not just being explained by the latent classes. I am getting more with the latent classes and as evidence of that even schools with a high percent of low SES students have a sizeable group of good performers. For instance, there are schools with 77%, 89% and 92% of students receiving FRL, nonetheless these very same schools have 36%, 34% and 31% of students who are HPs. In addition, these very same schools have 34%, 25%, and 39% of students who are LPs. If you remember, LPs have an initial status at the proficient level on average. Despite the fact that these students are low SES, there is variability in their performances captured by LCs. For this very reason, the continued research on relevant variability that may affect the aggregate accountability index needs to be pursued.

In addition, I suggest studying MLGMM based effects of different growth profile parameters, including the shape and rate of growth and the number of growth profiles, could also strengthen the estimation, applicability, and interpretability of cluster effects that are estimated with MLGMM. Particularly when including populations whose growth may be better represented with a non-linear function, such as LEPs, in their initial stages of language acquisition (i.e., less than four years). At some point, estimation of change in school effect will become a very important topic, possibly supporting the proficient/not-proficient classification based on VAM estimates, as long as the bias is controlled and is no longer differential depending on whether the school is stronger or weaker.

I now focus specifically on the actual interpretations and uses of data by professionals in educational contexts: teachers, school leaders, policy-makers and other stakeholders to expand validity discussion. Actual interpretations and uses are invariably shaped by local users' purposes, frequently require attention to multiple sources of evidence about students' learning

and the factors that shape it, and depend on local capacity to use such information well (Moss, 2016). There is a distinction between instrumental and conceptual uses of test based information. An instrumental approach entails 'using the results to make decisions [based] directly on test data without considering why test scores are low and a conceptual approach entails identifying patterns followed by systematic exploration of possible explanations, [which] requires collection and examination of other data (Murnane, Sharkey, and Boudett, 2009). Validity theory in educational measurement tends to support instrumental approaches based on a priori decision rules.

Moss (2016) argues that we also need validity theory to support 'conceptual' approaches that help educators connect the test-based data to their own practice and to consider explanations and explore solutions. It is here that the primary potential of testing to improve schooling lies. While the validity of such conceptual uses is ultimately a local responsibility (i.e., a school's district research and evaluation department, if it exists). Empirical studies of test use by teachers, administrators and policy-makers show that actual interpretations and uses of test scores in context are shaped by local capacity to use such information well (i.e., small school districts in North Carolina do not have a research and evaluation department) (Coburn & Turner, 2012). Using granular data analysis for sound decision making requires a more complex theory of validity that can shift focus as needed from the intended interpretations and uses of test scores that guide test developers to local capacity to support the actual interpretations, decisions and actions that routinely serve local users' purposes. While actual uses can be instrumental, more often than not a conceptual approach is needed to support educators in connecting test-based information to their practice to explain outcomes, frame questions or problems and explore solutions (Coburn &Turner, 2012).

Moss (2016) proposes that a detailed understanding of the whole educational system is needed in order to facilitate the conceptual use of data. Her goal is to use these and other data not just to identify problems, but, equally important, to develop explanations and to explore possible solutions. For instance, the practice of teachers is different from the practice of school leaders or policymakers or measurement specialists. And, professional role is only one of the many ways in which context matters. As Brown and Duguid (2000a) and Coburn and Talbert (2006) suggested, a coherent systemic strategy for evidence based practice may require a system of evidence use that allows for and supports access to different kinds of evidence for different purposes at different levels of the system. Individuals with different work roles have substantively different data needs. For instance, a curriculum director maybe interested in which program is working more effectively (greater growth with a given treatment), teachers may be interested in why some specific individuals in their class are not responding as well as other students similar to them, the superintendent and the school board may be interested in effective use of funding, and a defensible accountability system that supports the overall district goals and strategies (i.e., student learning).

In conclusion, as several scholars propose, a strategy for evidence-based reform must acknowledge these differences and create mechanisms to bring productive dialogue and coordination across the different levels and functions. For this purpose, long-range research is needed that pursues multiple validity arguments, which represent individual and aggregate score interpretations and uses to contribute to conceptual use. In return, information gained can guide particular inquiries into organizational capacity to use data well and ultimately contribute to the value of tests for enhancing education practice (Chalhoub-Deville, 2016; Moss, 2016; Coburn & Turner, 2012; Moss, 2013). For instance, this study only included LEPs that had been in the system receiving ESL services for at least four years by time zero. It was observed that there was

enough variability (on reading EOGs) in this population to be captured by the four latent classes. We can assume that the test was capturing variability of different achievement levels of the intended content for the LEP subpopulation.

However, if LEPs receiving less than four years of ESL services would have been included, then more than likely these students would have been classified as PLPs by the model because EOGs are not measuring only content but also language proficiency for this group. With respect to validating the use of educational tests for accountability purposes, future research should investigate the role that including all LEPs is playing on the estimate of the value-added scores. Particularly now that all LEPS (except first year LEPs) must be included in the value-added scores computation when EOGs are not measuring the intended content for LEPs. Such an evidence needs to be provided because this test might not be suitable to measure school effectiveness for LEPs in early stages of language development. After all, a fundamental assumption in the use of students' test performance to evaluate schools is that the curriculum, instruction and assessment are well aligned. Studies should consider how including LEPs who are at the early stages of acquiring the English language may affect the aggregate level scores (i.e., value-added score estimates).

Following the recommendation above of trying to develop explanations, I suggest the continued exploration of MLGMM for student-level and aggregate-level analysis . The research of growth trajectories studies in conjunction with relevant individual variables can explain why some individuals (or group of individuals with certain characteristics) never achieve reading proficiency and how some specific learning programs can help these students who are at risk from negative educational outcomes (e.g., permanently not being proficient, school drop, and others). For instance, from all the students in this study, 53% of students were reading proficient by time 3. From this group, 37% are HPs, which means that these students started at a proficient

level and maintained it for the duration of this study; 2% of students were S which means that they started not proficient but achieved proficiency by time 3 and the rest; 14% of student were LPs who hovered just above the mean (i.e., proficient) and maintained their proficiency status till the end.

Remember, the mean of this study was scaled to each year's average and it is also the point from which students move from proficient to not proficient status. The remaining 47% of students were not proficient at the beginning of this study and remained like that for the rest of the study. From this group 31% are PLPs who started not proficient and maintained this status till time 3. Meanwhile, the other 16% of the students were LPs that hovered just below the mean and maintained the not proficient status all along the four years of this study. Also for a more complete understanding, future research should include other less studied subpopulations (such as LEPs) interacting with other individual variables such as ethnicity and SES nested within an array of remedial programs. This study found that LEP did not predict the membership likelihood for PLPs or LPs. However, receiving FRL predicted the membership likelihood for PLPs and S. In this study most LEPs were Hispanic and most of them also received FRL. This study's results may not be generally applied to school districts, with different LEPs' SES status make up.

This study also shows that some subgroups such as low SES students and LEPs have lower average initial scores at school and that they need a much faster growth rate to acquire proficiency in reading (a student with a lower initial status will need a stronger growth rate) . The school system's average growth rate will not be sufficient to help these students achieve success. More in depth information is needed to understand what would stimulate a healthy growth rate in these subpopulations. For instance, information drawn from different areas (i.e., curriculum, psychology, language acquisition, behavioral management) is needed to understand why implemented remedial programs do not help to improve these subpopulation's educational goals.

Also in alignment with trying to understand all levels of the educational system, investigations should be conducted on whether the funding for remedial programs continue long enough to observe an adequate growth rate, or whether the funding was adequate every year (per number of students), or whether the remedial program or programs were appropriately implemented, or even a more basic question of whether school officials know the growth rate size per year so that a failing student (from a given initial status) can reach proficiency, let's say, in three years. There are so many organizational level variables that are unknown but relevant that could affect student learning. It is only with a finer grained understanding of why students are consistently failing aided with methodological approaches (such as longitudinal analysis combined with multilevel models and GMMs) that we identify problems and explore solutions. In summary, I support the continued exploration of MLGMM for sound decision-making in educational contexts because using deficient data (such as the evaluation of programs based only on cross-sectional data i.e., percent of students who are proficient, or only using two data points in time for the evaluation of programs) is not only insufficient but it also equates to professional malpractice.

Chalhoub-Deville (2016) adds to the validity discussion by arguing that validity claims also entail the whole educational system (e.g., policy makers, test developers, school leaders) performance including the allocation of roles and responsibilities among school officials, policy makers, test developers among others. She argues that what is lacking in consequential research are structures that make explicit the interconnections among policy stipulations, testing capabilities, and those impacted – at the individual, group, and societal levels through allocation of roles and responsibilities. This argument also aligns with Theory of Action (TOA) mandate of research into consequences at various levels. TOA demands to move validation beyond individual score interpretation and use, and it expands it to assess the overall effectiveness of a

system for a more complete accountability model (Chalhoub-Deville, 2016). For instance, there are some school districts in California and Illinois that have moved to attach responsibility to school officials for the strategies (for students' growth) they develop and put into action. The logic of attaching accountability to school officials for the decisions they make is based on the fact that it should be imperative that they should make informed decisions based on their expertise in the curriculum area they lead thereby minimizing the cost of those decisions to other stakeholders (i.e., teachers).

Additionally, Chalhoub-Deville (2016) argues that the scholarship needed to engage in defining roles and allocating responsibility for consequential research at the policy development phase is practically nonexistent in the field. Future research should investigate the impact of assessing schools mainly on cross sectional information such as percent proficient instead of using more appropriate methodologies that better capture student learning over time such as growth models. This policy creates incentives for school leaders to stop any further inquiries when they consume only cross-sectional data. School leaders do not have any motivation to try to understand why some students are failing, or which programs work best under which contexts, or whether value-added scores based only on reading scores are sufficient to make decisions . As it is, percent proficient accountability indices disproportionately penalizes low SES schools and unfairly reward high SES schools. In short, percent proficient negatively impacts schools that house students who indeed need district resources the most because the current accountability classification of school effect on students performance bares very little validity. Policy makers seem to be in denial, thinking that a poor accountability index somehow will trickle down good judgment and decisions throughout all the layers of educational leadership (i.e., district superintendents, assistant superintendents, curriculum directors, principals and teachers). One can only speculate why policy makers show no interest in paying attention to the evidence and

developing a well-informed comprehensive policy that actually supports a strategy to improve students' learning. Incentives from the top must be put in place to redirect behavior down the pipeline. In order to rectify this state of affairs, metrics based on growth need to be put in place. Another topic lacking at the policy level is the impact of teachers low wages in North Carolina (and if this issue impacts high teacher turnover, particularly in very poor schools).

Teachers in NC are one of worst compensated in the country, and they are not compensated for additional training. Schools with a large number of low performing students need very well trained teachers to cope with several challenges that their students bring with them (poverty, LEPs, students with disabilities, and others). In this particular school district, we are talking about 60% of schools with more than 50% of students receiving FRL.

However, it is hard to believe that policy makers can think that by paying teachers less, more effort can be demanded from them. For this school district, the poorer the school is the larger the teacher turnover rate (more than 20%), and the richer the school is the lower teacher turnover is (less than 10%). It also worth mention that this very same poorer school received a failing grade from the state while these same richer schools received a glowing review.

If we put into dollars the cost (of losing teachers) due to the fact that this district is using an invalid accountability metric (because it disproportionately negatively impacts low SES schools) vs. using a more informed metric, that amount would surmount to about two million dollars just for this district. Because a less informed accountability metric classifies twice the number of schools as failing (compared to a more informed metric). For all the reasons explained above a comprehensive understanding of the whole educational system is needed, policy theories, research procedures, and communication systems, to investigate and address potential (intended/unintended, positive/negative) consequences ultimately on student learning. Every

time a student does not receive the appropriate remedial course of action, for the adequate amount time under the right implementation, that child is paying the cost of the system's inefficiencies.

REFERENCES

Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. Springer: New York, NY.

Asparouhov, T. & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing, Inc.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*, London: Arnold.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods, 9*, 3-29.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. New York, NY: John Wiley.

Boscardin, C., Muthén, B., Francis, D., & Baker, E. (2008). Early identification of reading difficulties using heterogeneous developmental trajectories. *Journal of Educational Psychology, 100,* 192-208.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345,370.

Braun, H. (2016). Meeting the challenges to measurement in an era of accountability. Meeting the Challenges to Measurement in an Era of Accountability, c, 1–426.

Brown, J. S., & Duguid, P. (2000a). Balancing act: How to capture knowledge without killing it. Harvard Business Review, 78.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park CA: Sage.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33*, 261-304.

Burstein, L., Linn, R. L., & Capell, F.J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics, 3*, 547-385.

Chalhoub-Deville, M. (October 01, 2016). Validity theory: Reform policies, accountability testing, and consequences. Language Testing, 33, 4, 453-472.

Chudowsky, N., Chudowsky, V., & Kober, N. (2007). *Answering the question that matters most: Has student achievement increased since No Child Left Behind?* Washington DC: Center on Educational Policy.

Clogg, C.C., & Goodman, L.A. (1985). Simultaneous latent structural analysis in several groups. In N. B. Tuma (Ed.), *Sociological Methodology* (pp. 81-110). San Francisco: Jossey-Bass Publishers.

Coburn, C. E., & Talbert, J. E. (2006). Conceptions of evidence use in school districts: Mapping the terrain. American Journal of Education, 112, 469–495.

Coburn, C. E., & Turner, E. O. (2012). The practice of data use: An introduction. American Journal of Education, 118, 99–111.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates.

Croudace, T. J., Jarvelin, M. R., Wadsworth, M. E., & Jones, P. B. (2003). Developmental typology of trajectories to nighttime bladder control: Epidemiologic application of longitudinal latent class analysis. *American Journal of Epidemiology, 157*, 834-842.

Dayton, C. M (1991). Educational applications of latent class analysis. *Measurement and Evaluation in Counseling and Development, 24,* 131-141.

Doran, H. C., & Lockwood, J. R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics, 31*, 205-230.

Duncan, T.E., Duncan, S.C., & Strycker, L.A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications (2nd ed.)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling, 15*, 75-95.

Feldman, B. J., Masyn, K. E.. & Conger, R. D. (2009). New approaches to studying problem behaviors: A comparison of methods for modeling longitudinal, categorical adolescent drinking data. *Developmental Psychology, 45*, 3, 652-676.

Garcia, E. (2015). Inequalities at the starting gate, cognitive and non-cognitive skills gaps between 2010- 2011 kindergarten classmates. Economic Policy Institute.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 215-231.

Haertel, E. (2013). How is testing supposed to improve schooling? Measurement: Interdisciplinary Research & Perspective, 11(1–2), 1–18.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 171- 196). Greenwood, CT: Information Age Publishing, Inc.

Henry, K., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling, 17,* 193-215.

Jackson, C.K. (2012). Teacher quality at the high school level: the importance of accounting for tracks. National Bureau of Economical Research.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27,* 385-409.

Kane, M. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (Vol. 4, pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage Publications.

Kreft, I. G. G., de Leeuw, J, & Aiken, L. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1-22.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group- based approaches using growth mixture modeling. *Journal of Quantitative Criminology, 24,* 1-31.

Kreuter, F., Yan, T., & Tourangeau, R. (2008). Good item or bad – can latent class analysis tell?: The utility of latent class analysis for the evaluation of survey questions. *Journal of the Royal Statistical Society, Series A, 171,* 723-738.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lazarus, S. S., Wu, Y., Altman, J., & Thurlow, M. L. (2010). The characteristics of low performing students on large-scale assessments. *NCEO brief*. Minneapolis: National Center on Educational Outcomes, University of Minnesota.

Lindley, D. V., & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B, 34,* 1-41.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievement. *Electronic Journal of Statistics, 1,* 223-252.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499-532.

Marsh, H. W., Ludtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling errors. *Multivariate Behavioral Research, 44,* 764-802.

Messick, S. J. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13–103). New York, NY: American Council on Education/Macmillan.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley & Sons.

McQuarrie, A. D .R., & Tsai, C. L. (1998). *Regression and time series model selection*. World Scientific, London, UK.

Miles, J., & Shevlin, M. (2000). *Applying regression and correlation: A guide for students and researchers.* Thousand Oaks, CA: Sage.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. Journal of Educational Measurement, 50, 91–98.

Moss, P. A. (2016). Shifting the focus of validity for test use. Assessment in Education: Principles, Policy & Practice, 23(October), 1–16.

Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2009). Using student-assessment results to improve instruction: Lessons from a workshop. Journal of Education for Students Placed at Risk (JESPAR), 10, 269–280.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557-585.

Muthén, B. (1991). Analysis of longitudinal data using latent variable models with varying parameters. In L. Collins & J. Horn (eds.), *Best methods for the analysis of change. Recent advances, unanswered questions, future directions* (pp. 1-17). Washington DC: American Psychological Association.

Muthén, B. (2000). Methodological issues in random coefficient growth modeling using a latent variable framework: Applications to the development of heavy drinking. In J. Rose, L. Chassin, C. Presson & J. Sherman (Eds.). *Multivariate* a*pplications in substance use research* (pp. 113-140). Hillsdale, NJ: Erlbaum.

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.

Muthén, B. (2006). The potential of growth mixture modeling. *Infant and Child Development, 15*, 623- 625.

Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A, 172*, 639-657.

Muthén, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S.T ., Yang, C. C., Wang, C. P., Kellam, S., Carlin, J. & Liao, J. (2002). General growth mixture modeling for randomized preventive interventions. *Biostatistics, 3*, 459-475.

Muthén, B., Khoo, S.T., Francis, D. & Kim Boscardin, C. (2003). Analysis of reading skills development from Kindergarten through first grade: An application of growth mixture modeling to sequential processes. In S. R. Reise & N. Duan (Eds). *Multilevel Modeling: Methodological Advances, Issues, and Applications* (pp.71-89). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B. O., & Muthén, L. K. (2000). The development of heavy drinking and alcohol related problems from ages 18 to 37 in a U.S. national sample. *Journal of Studies on Alcohol, 61*, 290-300.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-469.

Muthén, L. K., & Muthén, B. O. (2016). *Mplus user's guide (V6.1).* Los Angeles:

Muthén & Muthén. Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group based approach. *Psychological Methods, 4*, 139-157.
Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*, 327-362.

Nezlek, J. B., & Zyzniewski, L. E. (1998). Using hierarchical linear modeling to analyze group data. *Group Dynamics: Theory, Research, and Practice, 2*, 313-320. No Child Left Behind Act of 2001, 20 U.S.C. § 6161.

Palardy, G., & Vermunt, J. K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics, 35*, 532-565.

Pollack, B. N. (1998). Hierarchical linear modeling and the "unit of analysis" problem: A solution for analyzing responses of intact group members. *Group Dynamics, 2*, 299-312.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth models. Quantitative applications in the social sciences*, Thousand Oaks, CA: Sage.

Preacher, K., Zyphur, M. & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15*, 209-233.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association, 53*, 873-880.

Quandt, R. E., & Ramsey J. B. (1972). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association, 73*, 730-752.

Raudenbush, S. W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.).* Newbury Park, CA: Sage Publications.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research Center. SAS Institute. (2008-2010). SAS, release 9.3 [Computer software]. Cary, NC: Author.

Schaeffer, C. M., Petras, H., Ialongo, N., Masyn, K. E., Hubbard, S., Poduska, J., & Sheppard, K. (2006). A comparison of girl's and boy's aggressive-disruptive behavior trajectories across elementary school: Prediction to young adult antisocial outcomes. *Journal of Consulting and Clinical Psychology, 74*, 500-510.

Schochet, P. Z., & Chiang, H. S. (2010). Error rates in measuring teacher and school performance based on student test score gains. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schochet, P. Z., & Chiang, H. S. (2013). What are the error rates for classifying teacher and school performance using value-added models? *Journal of Educational and Behavioral Statistics*, 38, 142-177.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464.

Shepard, L. A. (1993). Evaluating test validity. Review of Research in Education, 19, 405–450.

Singer, J. D. (1999). Using SAS Proc Mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323-355.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: multilevel, longitudinal and structural equation models*. London, England: Chapman & Hall/CRC.

Spillane, J. P. (2012). Data in practice: Conceptualizing the data based decision-making phenomena. American Journal of Education, 118, 113–141.

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching*. Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

Star, & J. A. Clausen (Eds.), *Measurement and prediction*. Princeton: Princeton University Press.

Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* Chichester, U.K.: John Wiley & Sons.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in a growth mixture model. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317– 341). Greenwich, CT: Information Age.

U.S. Department of Education (2006). Letter to Chief State School Officers: Assessment Requirements of NCLB and the Growth Model Pilot Project. http://www2.ed.gov/policy/elsec/guid/secletter/060221.html

U.S. Department of Education (2010). Race to the Top Program Executive Summary. http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random- effects population. *Journal of the American Statistical Association, 91*, 217-221.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J.A. Hagenaars & A.L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89-106). Cambridge, UK: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2008). *Latent GOLD 4.5 user's manual*. Belmont, MA: Statistical Innovations.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Wright, S. P., White, J. T., Snaders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models.* SAS Institute Inc. http://www.sas.com/resources/asset/SAS-EVAASStatistical-Models.pdf

Yumoto, F. (2011). Effects of unmodeled latent classes on multilevel growth mixture estimation in value-added modeling, 1-157.

APPENDIX A

SCHOOL VALUE-ADDED SCORES

Table A1a. Unstandardized Value-Added Scores without School SES

School	School SES	n	MLM VA	SD	MLGMM VA	SD
3	0.536	89	-0.487	1.048	-0.507	0.718
8	0.13	43	-0.384	0.655	-0.489	0.48
15	0.297	87	-0.523	0.747	-0.446	0.648
38	0.208	82	-0.487	0.736	-0.375	0.555
6	0.185	96	-0.439	0.65	-0.317	0.552
33	0.167	57	-0.572	0.766	-0.449	0.685
22	0.227	85	-0.29	0.802	-0.33	0.583
41	0.153	49	-0.547	0.777	-0.463	0.685
4	0.338	82	-0.22	0.827	-0.255	0.606
32	0.482	112	-0.19	0.877	-0.252	0.698
21	0.236	102	-0.182	0.746	-0.207	0.637
30	0.314	79	-0.189	0.807	-0.233	0.72
27	0.493	57	-0.109	0.834	-0.237	0.637
28	0.977	79	-0.131	0.813	-0.189	0.626
16	0.768	109	-0.109	0.888	-0.155	0.665
25	0.588	65	-0.185	0.948	-0.194	0.848
5	0.411	99	-0.134	0.867	-0.127	0.634
37	0.506	116	-0.134	0.852	-0.128	0.75
35	0.46	116	-0.194	0.862	-0.121	0.752
40	0.662	108	0.031	0.985	-0.119	0.76
34	0.877	79	-0.094	0.771	-0.117	0.817
18	0.586	118	0.095	1.056	-0.081	0.765
23	0.883	48	0.131	0.672	-0.002	0.595
31	0.513	82	-0.036	0.852	0.007	0.816
39	0.77	93	0.032	0.861	-0.008	0.644
12	0.923	67	-0.021	0.894	0.013	0.709
11	0.918	101	0.076	0.932	0.018	0.79
24	0.978	106	0.062	0.908	0.089	0.955
36	0.891	41	0.161	0.847	0.092	0.55
14	0.944	55	0.196	0.991	0.246	0.962
17	0.978	35	0.343	0.863	0.157	0.501
2	0.884	81	0.232	0.898	0.174	0.789
1	0.913	51	0.409	0.79	0.251	0.816
29	0.99	43	0.443	0.816	0.289	0.792
26	0.993	36	0.363	0.692	0.432	0.864
10	0.985	70	0.355	0.877	0.25	0.729
13	0.981	97	0.183	0.815	0.233	0.76
19	0.945	69	0.414	0.946	0.399	0.912
9	0.979	72	0.467	0.822	0.348	0.719
20	0.959	80	0.377	0.772	0.353	0.683
7	0.965	34	0.725	0.81	0.549	0.649

Table A1b. Unstandardized Value-Added Scores with School SES

School	School SES	n	MLM VA	SD	MLGMM VA	SD
7	0.965	35	-0.482	0.811	-0.357	0.911
9	0.979	72	-0.207	0.821	-0.068	0.754
1	0.913	52	-0.186	0.786	-0.059	0.816
21	0.236	102	-0.174	0.746	-0.160	0.579
19	0.945	69	-0.170	0.945	-0.167	0.918
29	0.99	43	-0.158	0.814	-0.003	0.767
18	0.586	118	-0.154	1.051	0.042	0.761
20	0.959	80	-0.127	0.773	-0.125	0.691
30	0.314	79	-0.091	0.809	-0.034	0.717
31	0.513	82	-0.080	0.849	-0.111	0.823
26	0.993	36	-0.080	0.692	-0.101	0.875
10	0.985	71	-0.076	0.875	0.036	0.724
22	0.227	85	-0.074	0.803	-0.029	0.582
17	0.978	35	-0.070	0.862	0.139	0.511
5	0.411	99	-0.067	0.862	-0.059	0.631
8	0.13	43	-0.064	0.655	0.080	0.483
4	0.338	82	-0.047	0.825	-0.012	0.622
2	0.884	81	-0.042	0.899	0.002	0.788
27	0.493	58	-0.035	0.826	0.104	0.633
40	0.662	109	-0.022	0.979	0.152	0.737
37	0.506	116	0.017	0.853	0.033	0.759
35	0.46	116	0.031	0.863	-0.027	0.759
36	0.891	41	0.036	0.843	0.077	0.543
6	0.185	96	0.039	0.649	-0.057	0.548
32	0.482	112	0.040	0.880	0.112	0.705
14	0.944	55	0.052	0.987	0.040	0.951
23	0.883	48	0.060	0.673	0.181	0.587
39	0.77	93	0.077	0.856	0.155	0.636
13	0.981	97	0.083	0.810	0.023	0.768
38	0.208	82	0.107	0.733	-0.008	0.600
41	0.153	49	0.121	0.779	0.053	0.660
25	0.588	67	0.140	0.942	0.096	0.919
33	0.167	57	0.161	0.766	0.067	0.693
11	0.918	103	0.177	0.945	0.219	0.774
16	0.768	109	0.206	0.887	0.246	0.645
15	0.297	87	0.222	0.751	0.157	0.653
24	0.978	108	0.231	0.910	0.145	0.986
12	0.923	67	0.240	0.890	0.181	0.708
34	0.877	79	0.276	0.770	0.297	0.826
3	0.536	89	0.385	1.070	0.392	0.720
28	0.977	79	0.391	0.812	0.424	0.627

Table A2a. Standardized Value-Added Scores without School SES

School	School SES	n	MLM VA	SE	MLGMM VA	SE
3	0.536	89	-3.961	0.111	-5.702	0.076
8	0.13	43	-3.374	0.100	-5.683	0.073
15	0.297	87	-5.944	0.080	-5.369	0.069
38	0.208	82	-5.414	0.081	-4.927	0.061
6	0.185	96	-5.909	0.066	-4.331	0.056
33	0.167	57	-5.174	0.101	-4.144	0.091
22	0.227	85	-2.793	0.087	-4.064	0.063
41	0.153	49	-4.505	0.111	-3.985	0.098
4	0.338	82	-1.894	0.091	-2.720	0.067
32	0.482	112	-1.726	0.083	-2.714	0.066
21	0.236	102	-1.828	0.074	-2.125	0.063
30	0.314	79	-1.564	0.091	-1.975	0.081
27	0.493	57	-0.561	0.110	-1.944	0.084
28	0.977	79	-0.918	0.091	-1.647	0.070
16	0.768	109	-0.729	0.085	-1.287	0.064
25	0.588	65	-1.174	0.118	-1.150	0.105
5	0.411	99	-0.998	0.087	-0.847	0.064
37	0.506	116	-1.100	0.079	-0.790	0.070
35	0.46	116	-1.837	0.080	-0.687	0.070
40	0.662	108	0.823	0.095	-0.629	0.073
34	0.877	79	-0.542	0.087	-0.479	0.092
18	0.586	118	1.461	0.097	-0.114	0.070
23	0.883	48	1.835	0.097	0.827	0.086
31	0.513	82	0.117	0.094	0.888	0.090
39	0.77	93	0.885	0.089	0.973	0.067
12	0.923	67	0.238	0.109	0.993	0.087
11	0.918	101	1.326	0.093	1.158	0.079
24	0.978	106	1.236	0.088	1.746	0.093
36	0.891	41	1.572	0.132	1.921	0.086
14	0.944	55	1.819	0.134	2.459	0.130
17	0.978	35	2.674	0.146	2.716	0.085
2	0.884	81	2.796	0.100	2.817	0.088
1	0.913	51	4.122	0.111	2.836	0.114
29	0.99	43	3.938	0.124	2.997	0.121
26	0.993	36	3.555	0.115	3.507	0.144
10	0.985	70	3.835	0.105	3.707	0.087
13	0.981	97	2.779	0.083	3.965	0.077
19	0.945	69	4.048	0.114	4.299	0.110
9	0.979	72	5.306	0.097	4.968	0.085
20	0.959	80	4.912	0.086	5.579	0.076
7	0.965	34	5.557	0.139	5.588	0.111

Table A2b. Standardized Value-Added Scores with School SES

School	School SES	n	MLM VA	SE	MLGMM VA	SE
7	0.965	35	-3.742	0.137	-2.702	0.154
21	0.236	102	-2.775	0.074	-3.820	0.057
9	0.979	72	-2.460	0.097	-1.429	0.089
1	0.913	52	-1.991	0.109	-1.043	0.113
18	0.586	118	-1.912	0.097	-0.243	0.070
20	0.959	80	-1.828	0.086	-2.382	0.077
19	0.945	69	-1.767	0.114	-2.045	0.111
29	0.99	43	-1.523	0.124	-0.530	0.117
30	0.314	79	-1.340	0.091	-1.153	0.081
22	0.227	85	-1.206	0.087	-1.394	0.063
31	0.513	82	-1.184	0.094	-1.870	0.091
5	0.411	99	-1.131	0.087	-1.861	0.063
10	0.985	71	-1.030	0.104	-0.268	0.086
26	0.993	36	-0.962	0.115	-1.097	0.146
8	0.13	43	-0.951	0.100	0.285	0.074
4	0.338	82	-0.856	0.091	-1.034	0.069
2	0.884	81	-0.731	0.100	-0.651	0.088
17	0.978	35	-0.693	0.146	0.926	0.086
27	0.493	58	-0.609	0.108	0.541	0.083
40	0.662	109	-0.565	0.094	1.317	0.071
37	0.506	116	-0.177	0.079	-0.369	0.070
35	0.46	116	0.000	0.080	-1.220	0.070
36	0.891	41	0.038	0.132	0.212	0.085
32	0.482	112	0.108	0.083	0.796	0.067
6	0.185	96	0.121	0.066	-2.074	0.056
14	0.944	55	0.158	0.133	-0.148	0.128
23	0.883	48	0.299	0.097	1.440	0.085
39	0.77	93	0.518	0.089	1.456	0.066
13	0.981	97	0.632	0.082	-0.462	0.078
41	0.153	49	0.809	0.111	-0.064	0.094
38	0.208	82	0.939	0.081	-1.011	0.066
25	0.588	67	0.947	0.115	0.330	0.112
33	0.167	57	1.281	0.101	0.087	0.092
11	0.918	103	1.568	0.093	2.098	0.076
12	0.923	67	1.922	0.109	1.410	0.086
16	0.768	109	2.060	0.085	3.027	0.062
24	0.978	108	2.284	0.088	0.906	0.095
15	0.297	87	2.372	0.081	1.400	0.070
34	0.877	79	2.828	0.087	2.561	0.093
3	0.536	89	3.121	0.113	4.363	0.076
28	0.977	79	3.941	0.091	5.174	0.071

Table A3. Estimates of Methodology Difference by School SES (high to low)

School	School SES	n	Methodology Difference	Status Change
8	0.13	43	0.609	
41	0.153	49	0.579	
33	0.167	57	0.561	
6	0.185	96	0.538	X
38	0.208	82	0.508	
22	0.227	85	0.484	
21	0.236	102	0.472	
15	0.297	87	0.393	X
30	0.314	79	0.372	
4	0.338	82	0.341	
5	0.411	99	0.247	
35	0.46	116	0.184	
32	0.482	112	0.155	
27	0.493	58	0.141	
37	0.506	116	0.124	
31	0.513	82	0.115	
3	0.536	89	0.086	
18	0.586	118	0.021	
25	0.588	67	0.019	
40	0.662	109	-0.077	
16	0.768	109	-0.213	
39	0.77	93	-0.216	
34	0.877	79	-0.354	
23	0.883	48	-0.361	
2	0.884	81	-0.363	
36	0.891	41	-0.372	
1	0.913	52	-0.400	
11	0.918	103	-0.406	X
12	0.923	67	-0.413	
14	0.944	55	-0.440	
19	0.945	69	-0.441	X
20	0.959	80	-0.459	X
7	0.965	35	-0.467	
28	0.977	79	-0.482	
24	0.978	108	-0.484	X
17	0.978	35	-0.484	
9	0.979	72	-0.485	X
13	0.981	97	-0.488	
10	0.985	71	-0.493	
29	0.99	43	-0.499	
26	0.993	36	-0.503	

Table A4. Classes Mixture Composition by School SES

School	School SES	n	Class PLP	Class HP	Class S	Class LP
8	0.13	43	27.27	45.45	0	27.27
41	0.153	49	9.43	58.49	0	32.08
33	0.167	57	11.67	56.67	1.67	30
6	0.185	96	9.71	56.31	0	33.98
38	0.208	82	11.36	55.68	1.14	31.82
22	0.227	85	24.44	46.67	0	28.89
21	0.236	102	20	53.33	0	26.67
15	0.297	87	17.58	57.14	1.1	24.18
30	0.314	79	21.59	46.59	0	31.82
4	0.338	82	26.14	47.73	1.14	25
5	0.411	99	22.64	46.23	0	31.13
35	0.46	116	18.9	39.37	2.36	39.37
32	0.482	112	30.51	38.98	0.85	29.66
27	0.493	58	31.67	33.33	0	35
37	0.506	116	23.14	38.02	2.48	36.36
31	0.513	82	18.82	34.12	1.18	45.88
3	0.536	89	31.91	45.74	1.06	21.28
18	0.586	118	40.63	33.59	0	25.78
25	0.588	67	26.47	42.65	2.94	27.94
40	0.662	109	41.59	26.55	0.88	30.97
16	0.768	109	29.27	35.77	0.81	34.15
39	0.77	93	32.99	32.99	3.09	30.93
34	0.877	79	37.65	24.71	2.35	35.29
23	0.883	48	42	26	0	32
2	0.884	81	40.23	28.74	3.45	27.59
36	0.891	41	40.91	34.09	0	25
1	0.913	52	44.64	19.64	1.79	33.93
11	0.918	103	42.34	27.03	4.5	26.13
12	0.923	67	29.58	30.99	0	39.44
14	0.944	55	40.35	29.82	3.51	26.32
19	0.945	69	42.86	20	5.71	31.43
20	0.959	80	40.96	21.69	2.41	34.94
7	0.965	35	47.22	13.89	5.56	33.33
28	0.977	79	36.47	24.71	0	38.82
17	0.978	35	48.57	17.14	0	34.29
24	0.978	108	36.94	21.62	6.31	35.14
9	0.979	72	53.85	15.38	2.56	28.21
13	0.981	97	33	26	2	39
10	0.985	71	48.61	22.22	0	29.17
29	0.99	43	45.65	17.39	2.17	34.78
26	0.993	36	35.14	10.81	5.41	48.65

Table A5a. School-Level Analysis: System Error Rates without School SES

Classification			Status Change from MLM to MLGMM		
Model	**Red**	**Blue**	**False Discovery Rate**	**False Non-Discovery Rate**	**Total Classification Change Rate**
Type I Error Rate Threshold Value=0.10					
MLM	12	13	2	4	6
MLGMM	13	14	0.05	0.10	0.15
Type I Error Rate Threshold Value=0.05					
MLM	8	11	-	4	4
MLGMM	11	12	-	0.10	0.10
Type I Error Rate Threshold Value=0.01					
MLM	8	10	-	3	3
MLGMM	10	11	-	0.07	0.07

Table A5b. School-Level Analysis: System Error Rates with School SES

Classification			Status Change from MLM to MLGMM		
Model	**Red**	**Blue**	**False Discovery Rate**	**False Non-Discovery Rate**	**Total Classification Change Rate**
Type I Error Rate Threshold Value=0.10					
MLM	7	7	6	4	10
MLGMM	7	5	0.15	0.10	0.25
Type I Error Rate Threshold Value=0.05					
MLM	3	6	3	4	7
MLGMM	5	5	0.07	0.10	0.17
Type I Error Rate Threshold Value=0.01					
MLM	2	3	2	1	3
MLGMM	1	3	0.05	0.02	0.07

Figure A1a. Unstandardized Value-Added Scores without School SES

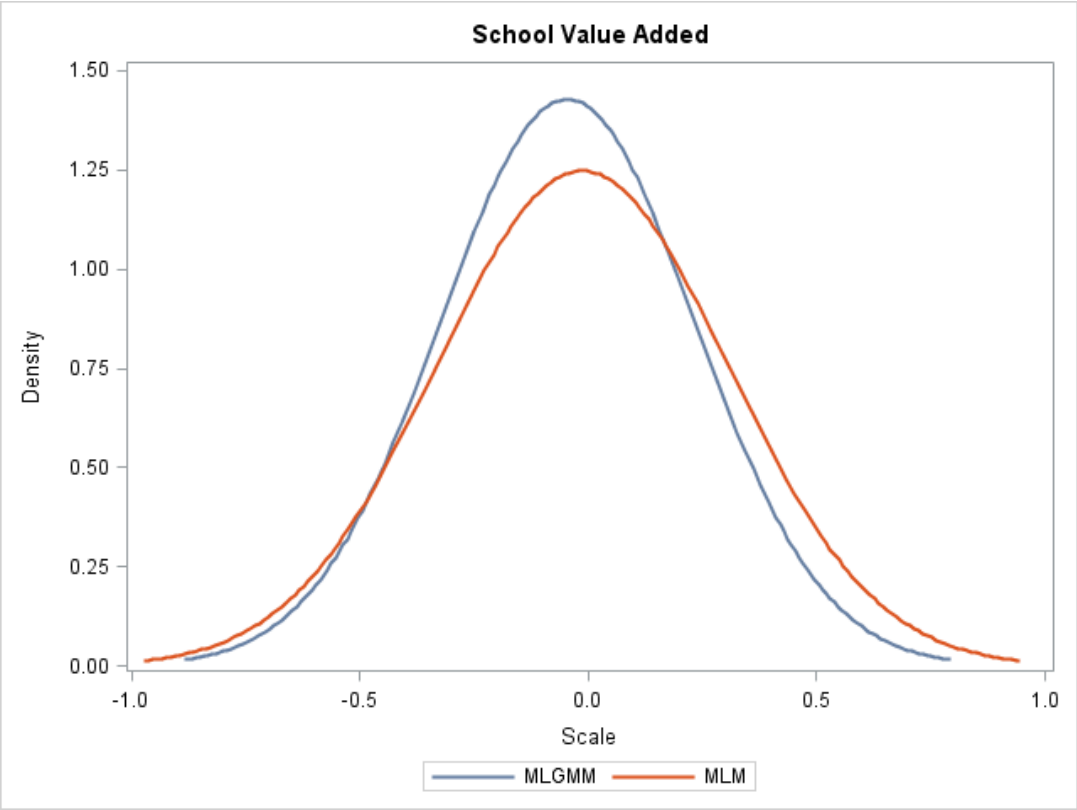


Figure A1b. Unstandardized Value-Added Scores with School SES

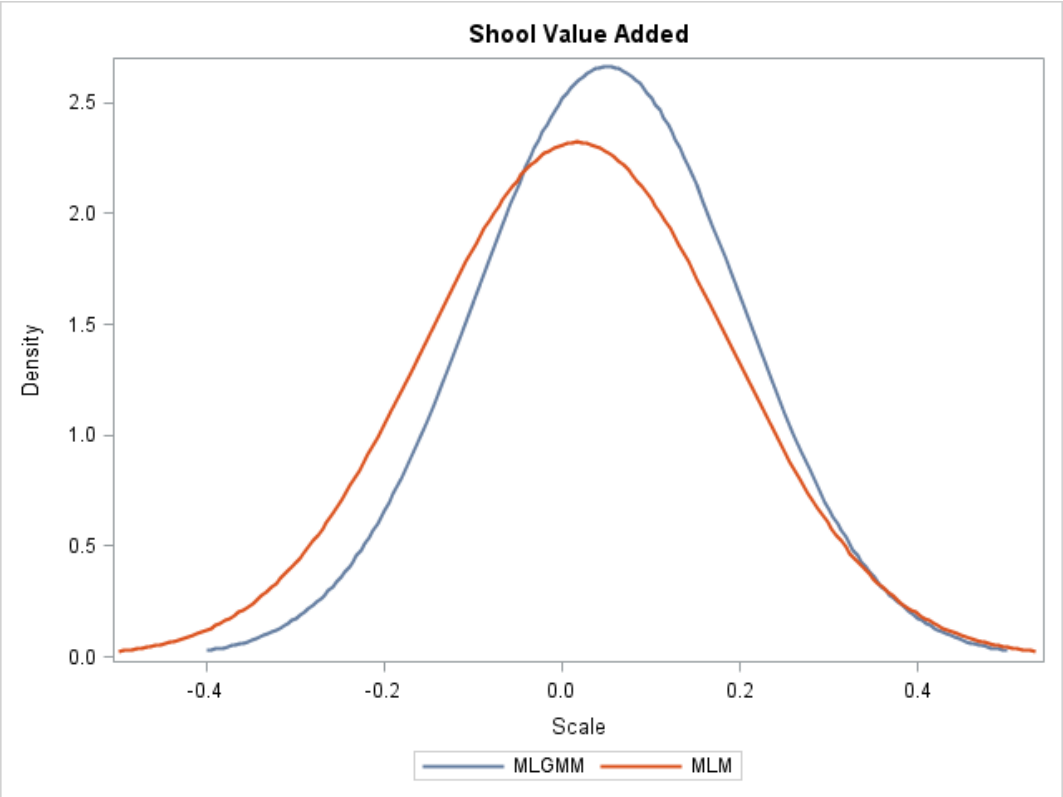


Figure A2a. Standardized Value-Added Scores without School SES

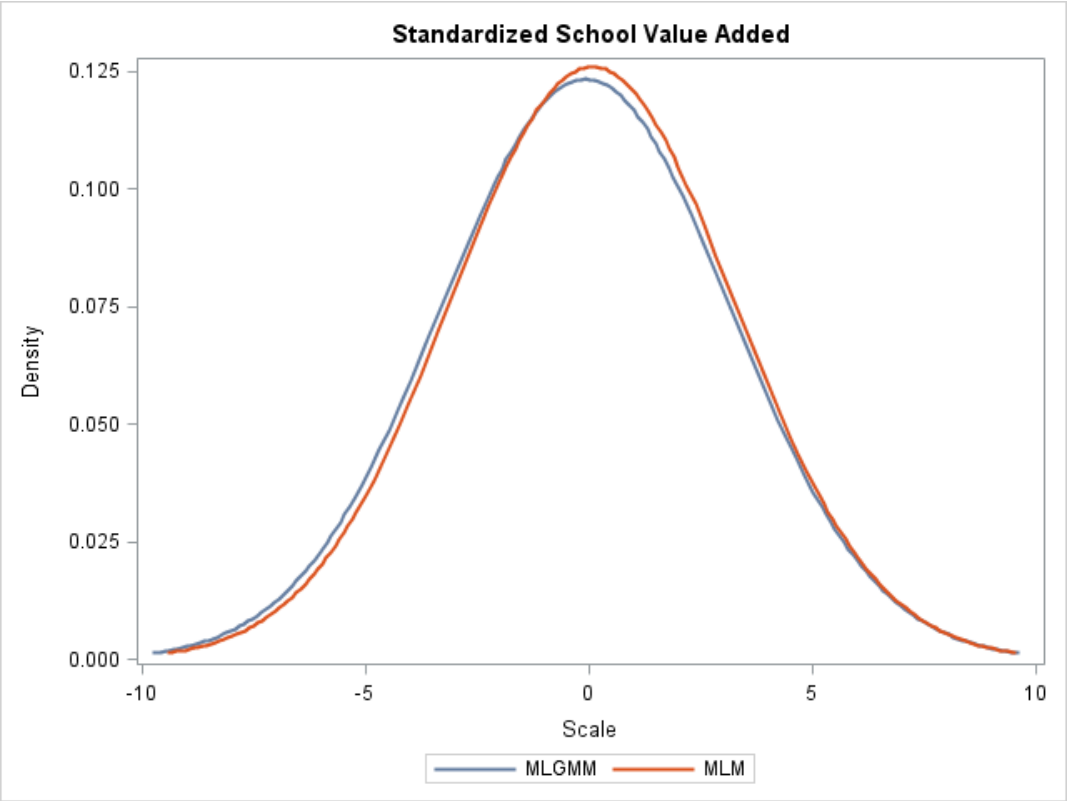


Figure A2b. Standardized Value-Added Scores with School SES

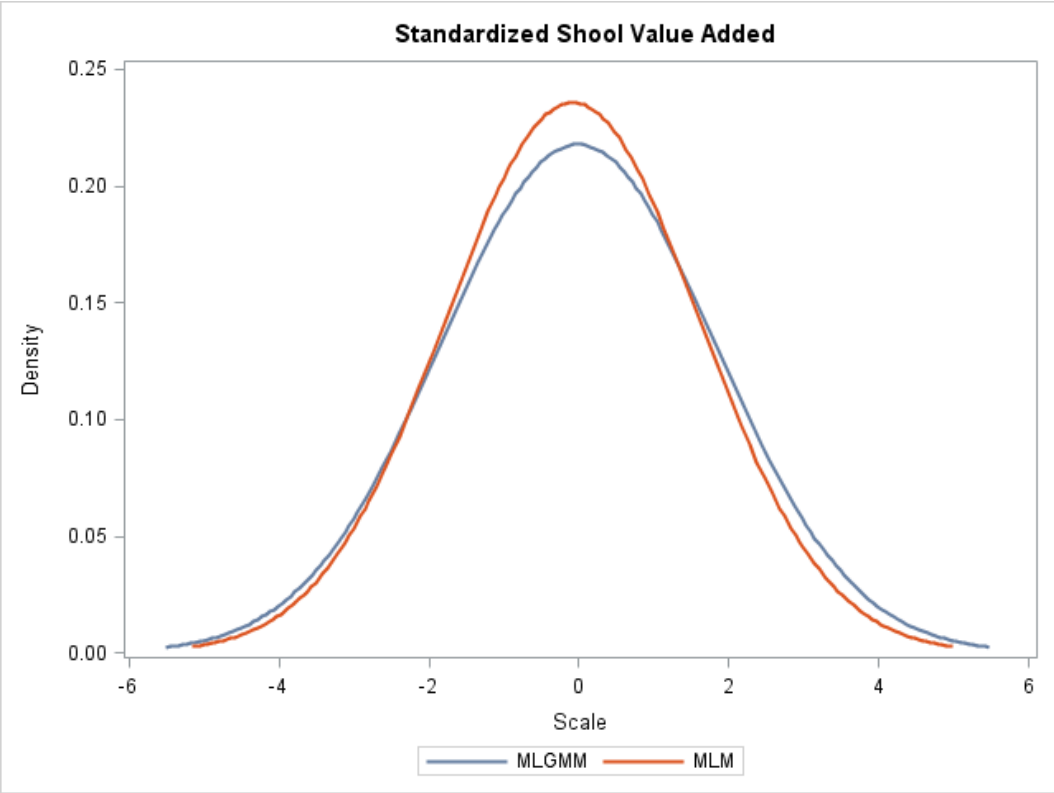


Figure A3a. Ranked Standardized Value-Added Scores without School SES

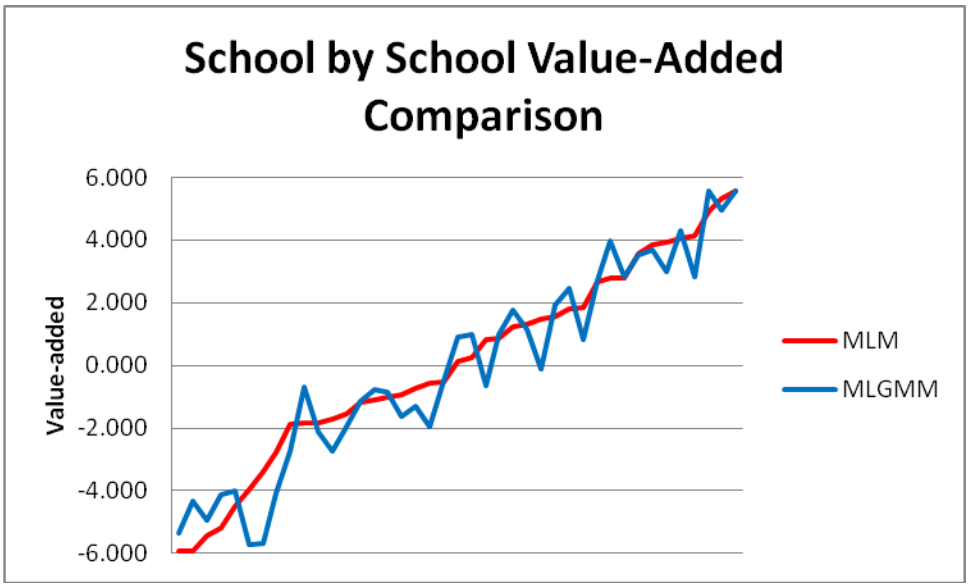


Figure A3b. Ranked Standardized Value-Added Scores with School SES

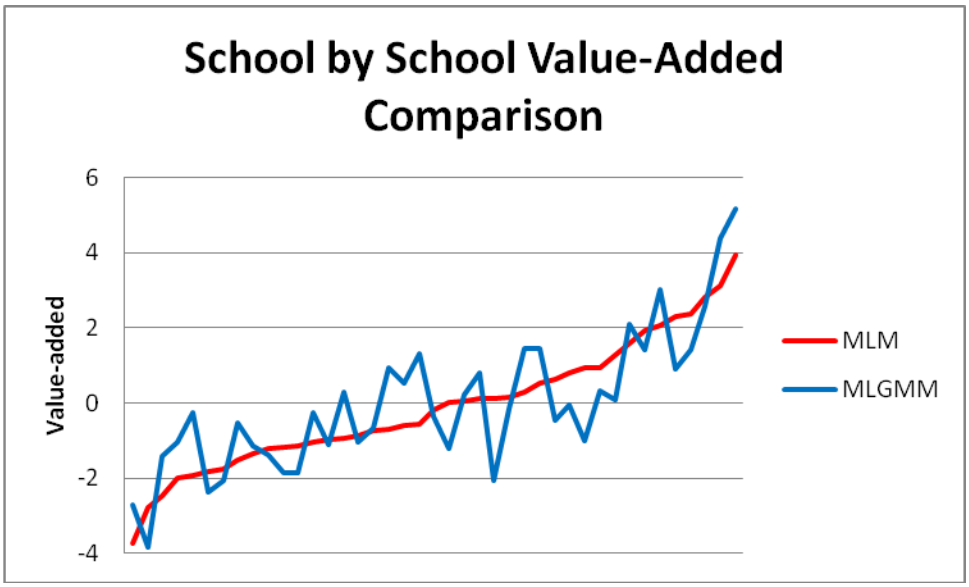


Figure A4a. Value-Added Scores SD Estimates without School SES

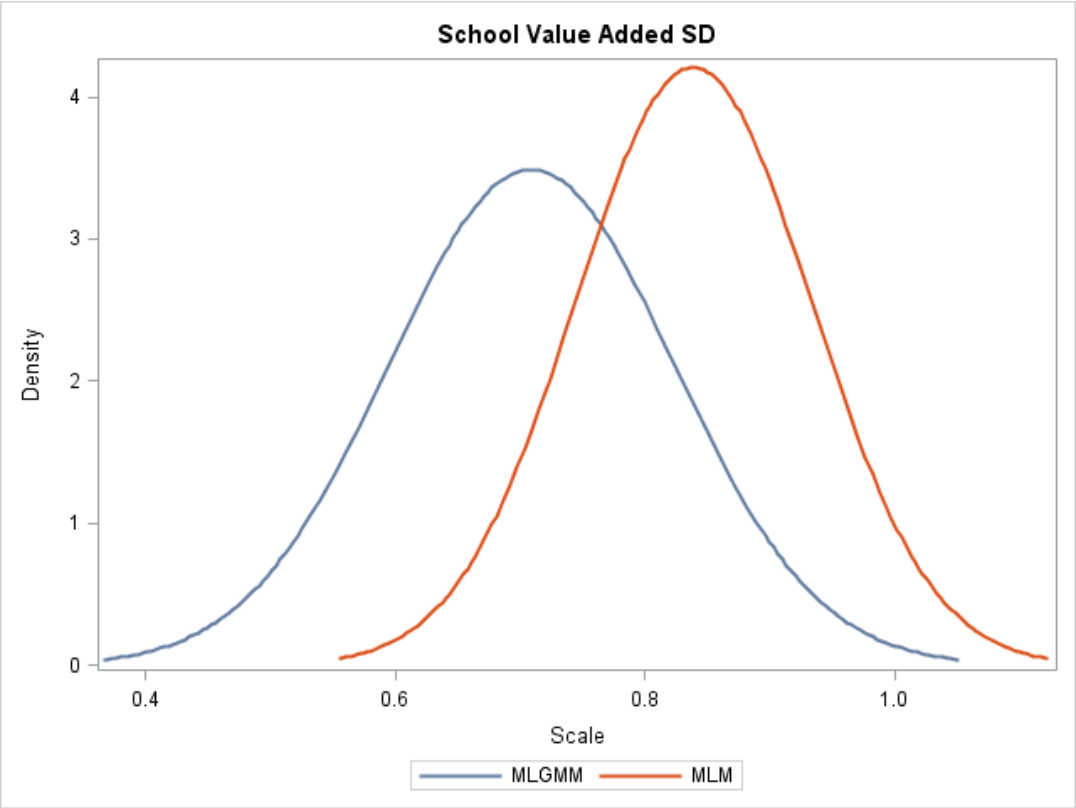


Figure A4b. Value-Added Scores SD Estimates with School SES

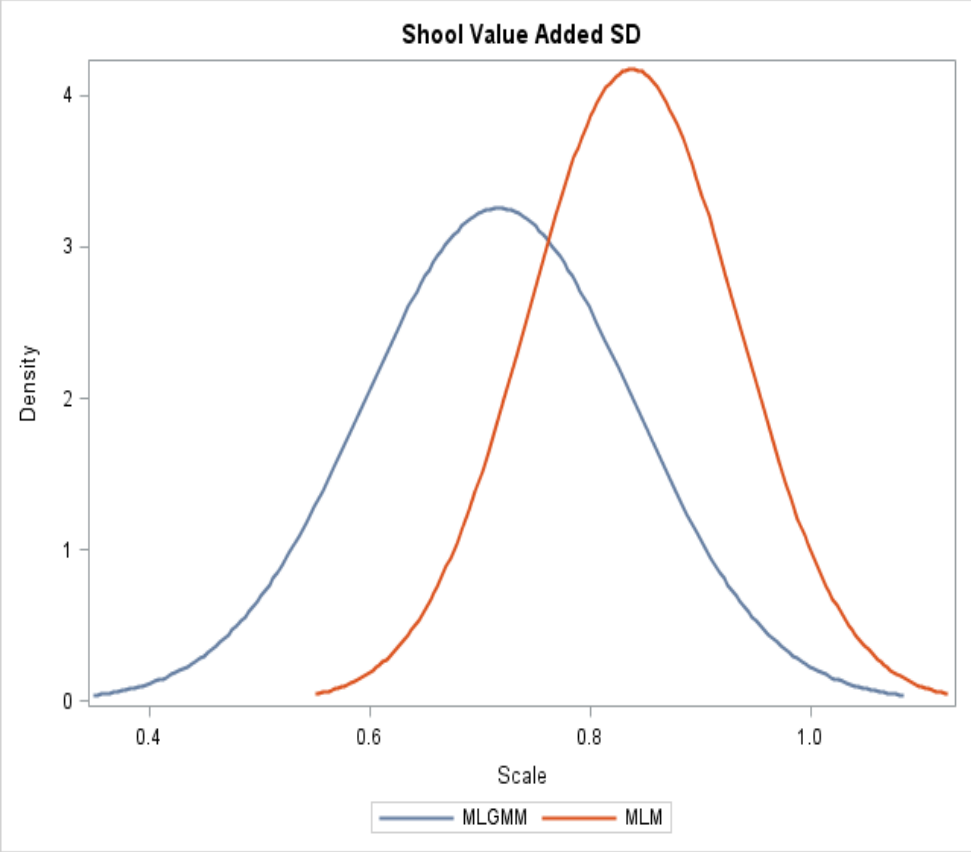


Figure A5a. Value-Added Scores SE Estimates without School SES

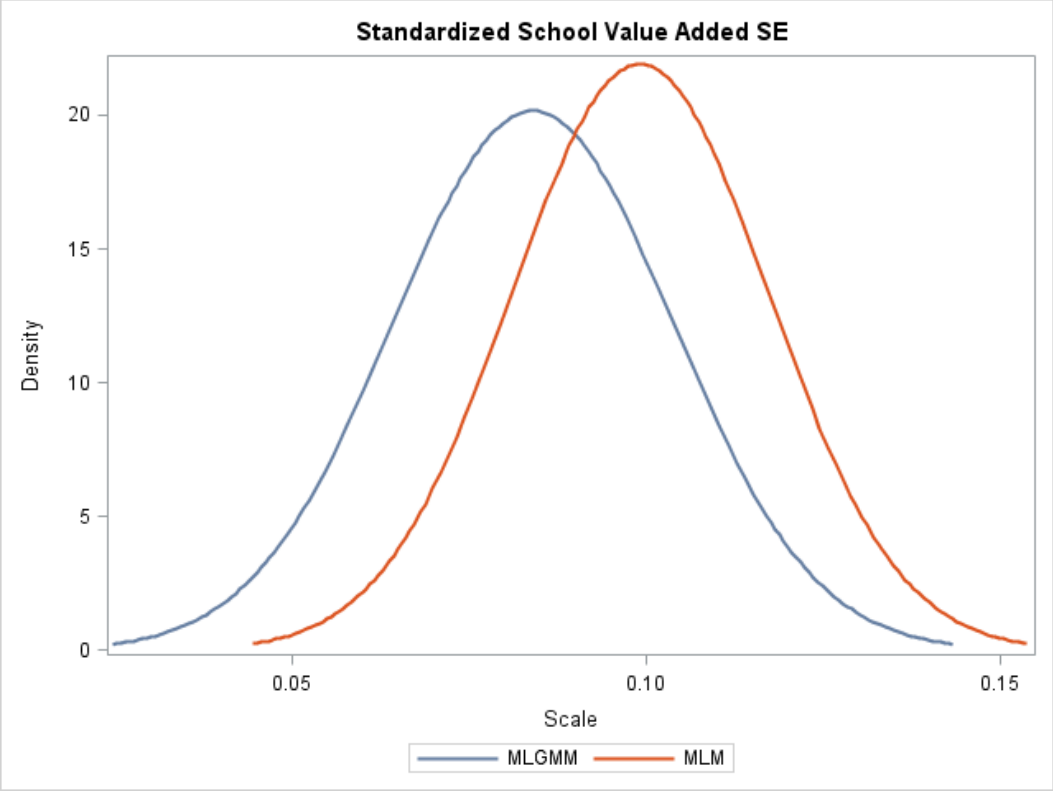


Figure A5b. Value-Added Scores SE Estimates with School SES

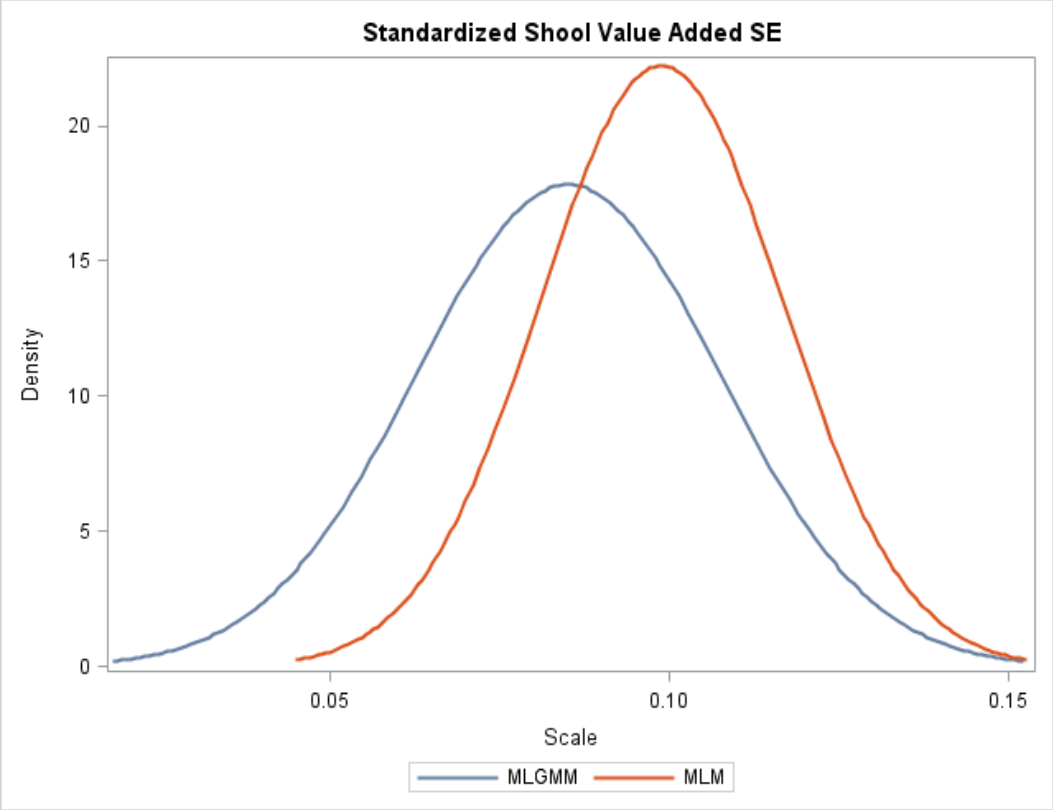


Figure A6a. Ranked Value-Added SE without School SES

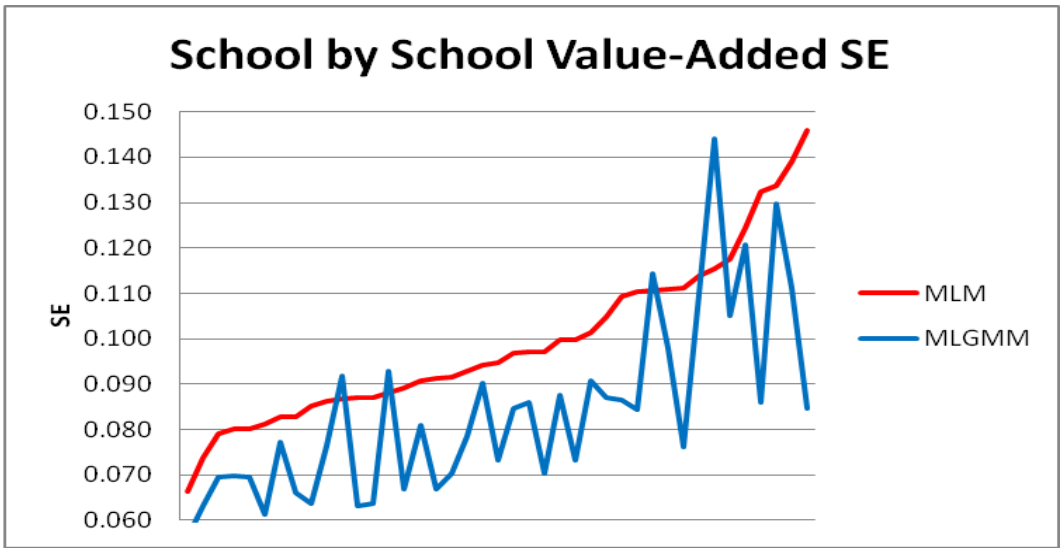


Figure A6b. Ranked Value-Added SE with School SES

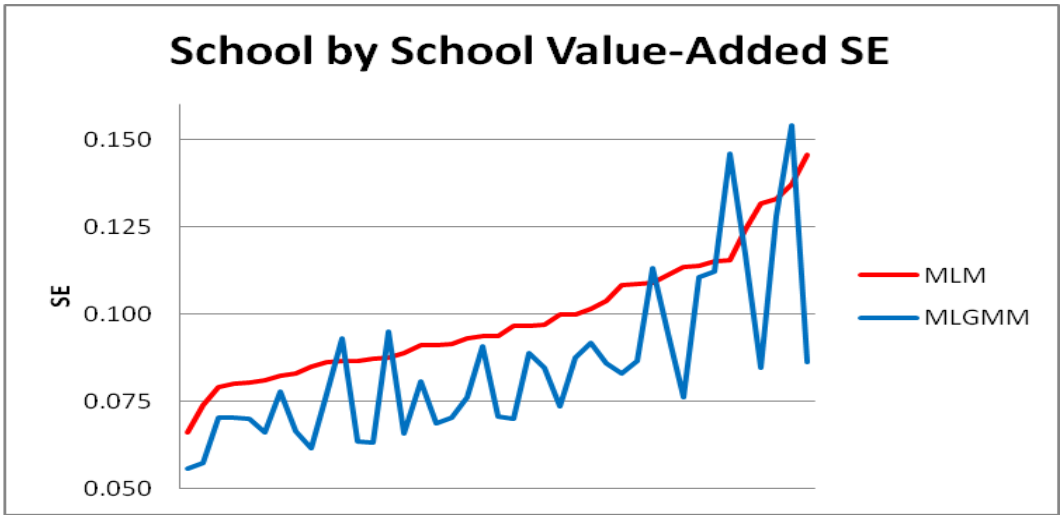


Figure A7. Standardized Value-Added Scores for all Models

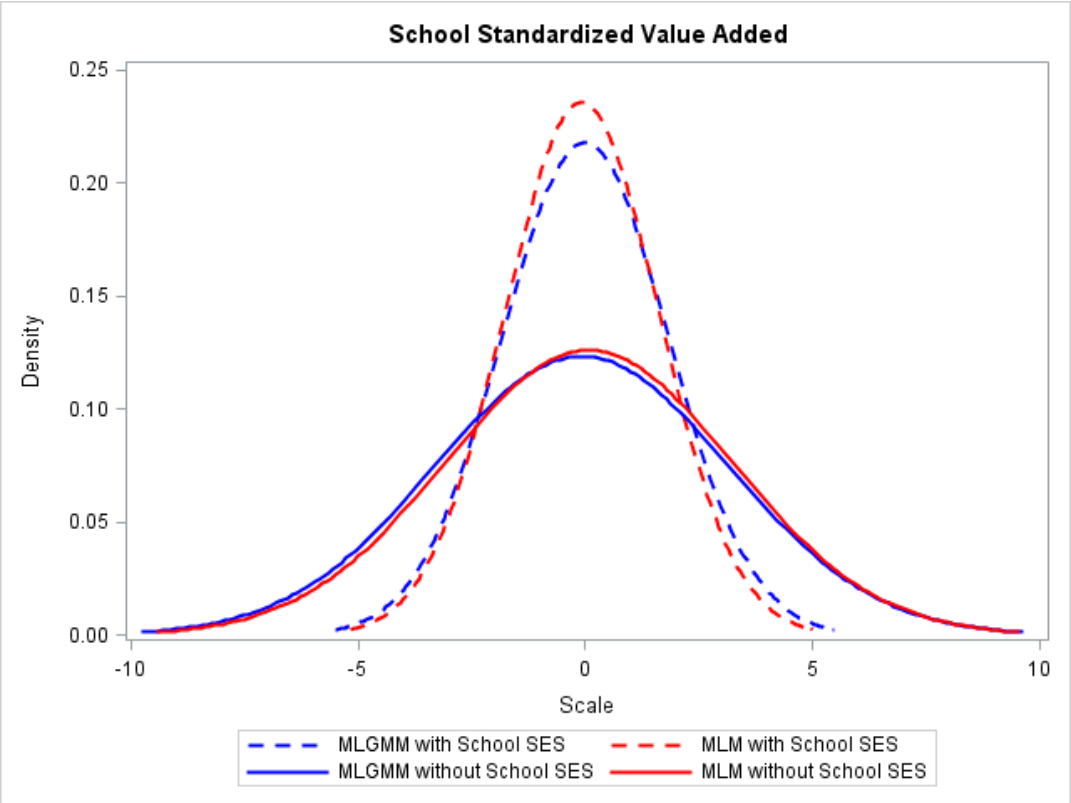


Figure A8. Standardized Value-Added Scores SE for all Models